Journal of Multivariate Analysis 100 (2009) 175–194

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Learning from dependent observations

Ingo Steinwart*, Don Hush, Clint Scovel

Information Sciences Group, CCS-3 MS B256, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

ARTICLE INFO

Article history:
Received 18 May 2006
Available online 12 April 2008

AMS subject classifications:
primary 68T05 (1985)
secondary 62G08 (2000)
62H30 (1973)
62M45 (2000)
68Q32 (2000)

Keywords:
Support vector machine
Consistency
Non-stationary mixing process
Classification
Regression

ABSTRACT

In most papers establishing consistency for learning algorithms it is assumed that the observations used for training are realizations of an i.i.d. process. In this paper we go far beyond this classical framework by showing that support vector machines (SVMs) only require that the data-generating process satisfies a certain law of large numbers. We then consider the learnability of SVMs for α-mixing (not necessarily stationary) processes for both classification and regression, where for the latter we explicitly allow unbounded noise.

Published by Elsevier Inc.

1. Introduction

In recent years Support Vector Machines (SVMs) have become one of the most widely used algorithms for classification and regression problems. Besides their good performance in practical applications they also enjoy a good theoretical justification in terms of both universal consistency (see [1–4]) and learning rates (see [5–9]) if the training samples come from an i.i.d. process. However, often this i.i.d. assumption cannot be strictly justified in real-world problems. For example, many machine learning applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes. Moreover, samples are often gathered from different sources and hence it seems unlikely that they are identically distributed. Although SVMs have no theoretical justification in such non-i.i.d. scenarios they are often applied successfully. One of the goals of this work is explain this success by establishing consistency results for SVMs under somewhat minimal assumptions on the data-generating process. Namely, we show that for any data-generating process that satisfies certain laws of large numbers there exists a sequence of regularization parameters such that the corresponding SVM is consistent. By general negative results (see [10]) on universal consistency for stationary ergodic processes this sequence of regularization parameters must depend on the stochastic properties of the data-generating process and cannot be adaptively chosen. However, we show that if the process satisfies certain mixing properties such as polynomially decaying α-mixing coefficients (see the definitions in the following sections) then a suitable regularization sequence can be chosen a priori. In addition, a side effect of our analysis is that it provides consistency for SVMs using Gaussian kernels even if the common compactness assumption of the input space is violated. Consequently, our consistency results for α-mixing processes generalize earlier consistency results of [1–3] with respect to both the compactness assumption on X and the i.i.d. assumption on the data-generating process.

* Corresponding author.
E-mail addresses: ingo@lanl.gov (I. Steinwart), dhush@lanl.gov (D. Hush), jcs@lanl.gov (C. Scovel).

0047-259X/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmva.2008.04.001

Relaxations of the independence assumption have been considered for quite a while in both the machine learning and the statistical literature. For example PAC-learning for stationary $\bar{\beta}$-mixing processes has been investigated in [11], and more recently, consistency of regularized boosting for classification was established for such processes. For a larger class of processes, namely α-mixing but not necessarily stationary processes, consistency of kernel density estimators was shown in [12]. For bounded, stationary processes with exponentially decaying $\bar{\alpha}$-mixing coefficients a consistent method for one-step-ahead prediction (also known as "static autoregressive forecasting", see [13]) was presented in [14]. Moreover, for this prediction problem [15] establishes consistency for a certain structural risk minimization approach under the assumption that the process is stationary and has polynomially decaying $\bar{\beta}$-mixing rates. For further results and references we refer the reader to [16,17].

Relaxations of the stationarity of the process are less common. In fact, to the best of our knowledge [12] is the only work which deals with such processes. One of the reasons for this lack of literature may be the fact that for non-identically distributed observations there is no obvious way to define a reasonable risk functional which resembles the idea of "average future error". On the other hand, it seems obvious that learning methods based on a modified empirical risk minimization procedure require at least that the process satisfies certain laws of large numbers. Interestingly, we will show that for processes satisfying such laws of large numbers there is always a "limit" distribution which can be used to define a reasonable risk functional. Moreover, for many interesting classes of processes the existence of such a limit distribution turns out to be equivalent to a law of large numbers.

The rest of this work is organized as follows: In Section 2 we will define the notions "laws of large numbers" and "limit" distributions for stochastic processes. We then discuss the relationship between these concepts and consider specific classes of stochastic processes that satisfy these definitions. We then recall some basic classes of loss functions and define consistency of learning algorithms for stochastic processes satisfying certain laws of large numbers. Finally, we show that SVMs can be made consistent for such processes. In Section 3 we then recall various mixing coefficients for stochastic processes. These coefficient are then used to establish consistency results for SVMs with a priori chosen regularization sequence. Finally, the proofs of our results can be found in Section 4.

2. Consistency for processes satisfying a law of large numbers

The aim of this section is to show that SVMs can be made consistent whenever the data-generating process satisfies a certain type of law of large numbers (LLNs). To this end we first recall some notions for stochastic processes and introduce these laws of large numbers in Section 2.1. In Section 2.2 we then recall some important notions for loss functions and risks. We also define consistency of learning algorithms for data-generating processes that satisfy a law of large numbers. Finally, we present and discuss our consistency results for SVMs in Section 2.3.

2.1. Law of large numbers for stochastic processes

In this subsection we mainly introduce laws of large numbers for general, not necessarily stationary stochastic processes. The concepts we will present seem to be quite natural and elementary, and therefore one would expect that they have already been introduced elsewhere. Surprisingly, however, we were not able to find any exposition that covers major parts of the material of this section, and thus we discuss the following notions in some detail.

Let us begin with some notations. Given a measurable space Z we write $\mathcal{L}_0(Z)$ for the set of all measurable functions $f : Z \to \mathbb{R}$, and $\mathcal{L}_\infty(Z)$ for the set of all bounded measurable functions $f : Z \to \mathbb{R}$. Moreover, for a set $B \subset Z$ we write $\mathbf{1}_B$ for its indicator function, i.e. $\mathbf{1}_B : Z \to \{0, 1\}$ with $\mathbf{1}_B(z) = 1$ if and only if $z \in B$. Let us now assume that we also have a probability space $(\Omega, \mathcal{A}, \mu)$ and a measurable map $T : \Omega \to Z$. Then $\sigma(T)$ denotes the smallest σ-algebra on Ω for which T is measurable. Moreover, μ_T denotes the T-image measure of μ, which is defined by $\mu_T(B) := \mu(T^{-1}(B))$, $B \subset Z$ measurable. In particular, if $\mathcal{Z} := (Z_i)_{i\geq 1}$ is a Z-valued *stochastic process* on $(\Omega, \mathcal{A}, \mu)$ then $\mu_{\mathcal{Z}}$ denotes the image measure of the map $\mathcal{Z} : \Omega \to Z^{\mathbb{N}}$. Furthermore, recall that $\mathcal{Z}$ is called *identically distributed* if $\mu_{Z_i} = \mu_{Z_j}$ for all $i, j \geq 1$, and *stationary in the wide sense* if $\mu_{(Z_{i_1+i}, Z_{i_2+i})} = \mu_{(Z_{i_1}, Z_{i_2})}$ for all $i_1, i_2, i \geq 1$. Finally, $\mathcal{Z}$ is said to be *stationary* if $\mu_{(Z_{i_1+i}, \ldots, Z_{i_n+i})} = \mu_{(Z_{i_1}, \ldots, Z_{i_n})}$ for all $n, i, i_1, \ldots, i_n \geq 1$.

As we will see later we are not interested in the data-generating process $\mathcal{Z}$ itself, but only in processes of the form $g \circ \mathcal{Z} := (g \circ Z_i)_{i\geq 1}$ for $g : Z \to Z'$ being measurable. In the following we call $g \circ \mathcal{Z}$ an *image* of the process $\mathcal{Z}$, and $\mathcal{Z}$ itself a *hidden* process. The following definition introduces laws of large numbers for stochastic processes by considering real-valued image processes:

Definition 2.1. Let $\mathcal{Z}$ be a Z-valued stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$. We say that $\mathcal{Z}$ satisfies the *weak law of large numbers for events (WLLNE)* if for all measurable $B \subset Z$ there exists a constant $c_B \in \mathbb{R}$ such that for all $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B \circ Z_i(\omega) - c_B\right| > \varepsilon\right\}\right) = 0. \tag{1}$$

Moreover, we say that $\mathcal{Z}$ satisfies the *strong law of large numbers for events (SLLNE)* if for all measurable $B \subset Z$ there exists a constant $c_B \in \mathbb{R}$ such that for μ-almost all $\omega \in \Omega$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_B \circ Z_i(\omega) = c_B. \tag{2}$$

It is obvious that $\mathcal{Z}$ satisfies the WLLNE if and only if the sequences $(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B \circ Z_i)$ converge in probability μ for all measurable $B \subset Z$. Consequently, the SLLNE implies the WLLNE but in general the converse implication does not hold. Moreover, if $\mathcal{Z}$ satisfies the WLLNE then the constants c_B in (1) must obviously satisfy $c_B \in [0, 1]$ for all measurable $B \subset Z$. Finally, if $\mathcal{Z}$ satisfies the WLLNE or SLLNE then every image $g \circ \mathcal{Z}$ also satisfies the WLLNE or SLLNE, respectively.

For i.i.d. processes the map $B \mapsto c_B$ clearly defines a probability measure on Z. Our next goal is to show that this remains true for general processes satisfying a WLLNE. To this end we first consider the averages $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\mu \mathbf{1}_B \circ Z_i$ of the probabilities of the event B:

Definition 2.2. We say that a Z-valued stochastic process $\mathcal{Z}$ on the probability space $(\Omega, \mathcal{A}, \mu)$ is *asymptotically mean stationary (AMS)* if for all measurable $B \subset Z$ the following limit exists

$$P(B) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu \mathbf{1}_B \circ Z_i. \tag{3}$$

The notion "asymptotically mean stationary" was first introduced for dynamical systems by Grey and Kieffer in [18]. We are unaware of any work that introduces this notion for general stochastic processes, though a similar idea already appears as assumption (S1) in [12].

Obviously every image of an AMS process is again AMS. Moreover, identically distributed – and hence stationary – processes are obviously AMS. In addition, for such processes $\mathcal{Z}$ we also have $P(B) = \mu_{Z_1}(B)$ for all measurable $B \subset Z$, and consequently, (3) defines a probability measure on Z. The following lemma whose proof can be found in Section 4 shows that the latter observation remains true for general AMS processes.

Lemma 2.3. *Let $\mathcal{Z}$ be a Z-valued AMS process on the probability space $(\Omega, \mathcal{A}, \mu)$. Then P defined by (3) is a probability measure on Z. We call P the* stationary mean of $(\mathcal{Z}, \mu)$.

It it is well known that not every stationary process satisfies a (weak, strong) law of large numbers for events. Consequently, we see that in general AMS processes do not satisfy a law of large numbers. However, the following theorem proved in Section 4 shows that the converse implication *is* true. In addition, it shows that the constants c_B in (1) define the stationary mean distribution.

Theorem 2.4. *Let $\mathcal{Z}$ be a Z-valued stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$ that satisfies the WLLNE. Then $\mathcal{Z}$ is AMS and the stationary mean P of $(\mathcal{Z}, \mu)$ satisfies*

$$\lim_{n\to\infty} \mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B \circ Z_i(\omega) - P(B)\right| > \varepsilon\right\}\right) = 0 \tag{4}$$

for all measurable $B \subset Z$ and all $\varepsilon > 0$. Moreover, if $\mathcal{Z}$ satisfies the SLLNE then μ-almost surely

$$\lim_{n\to 0} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_B \circ Z_i = P(B).$$

Eq. (4) shows that the stationary mean P describes with high probability our average observations from $\mathcal{Z}$. Given a loss function L (see Section 2.2 for definitions) it seems therefore natural to approximate the empirical L-risk of a function by the corresponding L-risk defined by P.[1] However, in order to make this ansatz rigorous we have to extend (4) to function classes larger than the set of indicator functions. We begin with the following result that shows that a law of large numbers for events implies a corresponding law of large numbers of bounded functions:

Lemma 2.5. *Let $\mathcal{Z}$ be a Z-valued stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$ that satisfies the WLLNE. Furthermore, let P be the asymptotic mean of $(\mathcal{Z}, \mu)$. Then for all $f \in \mathcal{L}_\infty(Z)$ we have*

[1] For i.i.d. observations one typically argues the other way around. However, for general stochastic processes the learning goal should be to minimize the future average loss. This loss is an empirical L-risk which can be approximated by the L-risk defined by P. In the training phase of empirical risk minimizers the latter L-risk is then approximated by the empirical L-risk of the already observed training samples. In this way P and the corresponding convergence rates in (3) and (4) tell us how well we can generalize from the past to the future.

$$\mathbb{E}_P f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i \tag{5}$$

in probability μ*, and*

$$\mathbb{E}_P f = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i. \tag{6}$$

Moreover, if $\mathcal{Z}$ *actually satisfies the SLLNE then the convergence in* (5) *holds* μ*-almost surely.*

For classification problems we usually can restrict our considerations to *bounded* functions, and hence Lemma 2.5 is all what we need. However, for regression problems with *unbounded* noise we have to consider *integrable* functions, instead. The following definition serves this purpose:

Definition 2.6. Let $\mathcal{Z}$ be a Z-valued AMS process on the probability space $(\Omega, \mathcal{A}, \mu)$ and P its asymptotic mean. We say that $\mathcal{Z}$ satisfies the *weak law of large numbers (WLLN)* if

$$\lim_{n\to\infty} \mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^{n} f \circ Z_i(\omega) - \mathbb{E}_P f\right| > \varepsilon\right\}\right) = 0 \tag{7}$$

for all $f \in L_1(P)$ and all $\varepsilon > 0$. Moreover, we say that $\mathcal{Z}$ satisfies the *strong law of large numbers (SLLN)* if for all $f \in L_1(P)$ we μ-almost surely have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f \circ Z_i = \mathbb{E}_P f. \tag{8}$$

Let us end this discussion by recalling some examples of types of stochastic processes that satisfy the above definitions.

Example 2.7 (*Independent Processes*). Obviously, i.i.d. processes satisfy the SLLN and this remains true for certain types of martingales. Moreover, by [19, Theorem 2.7.1] we see that a stochastic process $\mathcal{Z}$ for which all images $\mathbf{1}_B \circ \mathcal{Z}$ are independent the SLLNE is satisfied if and only if $\mathcal{Z}$ is AMS. Analogously, by Markov's inequality it is not hard to see that a stochastic process $\mathcal{Z}$ whose coordinates Z_i are pairwise independent satisfies the WLLNE if and only if $\mathcal{Z}$ is AMS.

Example 2.8 (*Ergodic Processes*). Recall that for invariant dynamical systems Birkhoff's theorem states that ergodicity is equivalent to the SLLN or SLLNE, and from this one can conclude that every stationary ergodic process $\mathcal{Z}$ satisfies the SLLN. In particular, if $\mathcal{Z}$ is an invariant ergodic dynamical system on $(\mathbb{R}^d, \mu)$ and $\mathcal{E}$ is an $\mathbb{R}^d$-valued i.i.d. process then the process $\mathcal{Z}+\mathcal{E}$ is stationary and ergodic and hence satisfies the SLLN. More information on ergodicity can be found in e.g. the books [20, 21].

Example 2.9 (*Markov Chains*). Stationary homogeneous Markov chains satisfying the "Doeblin condition" (see e.g. [22, p. 197] or [23, p. 156]) are known to satisfy the SLLN (see [22, p. 219]). Moreover, simple assumptions ensuring Doeblin's condition can be found in [22, p. 192f], and for conditions similar to Doeblin's condition we refer the reader to [23] and the references therein. In addition, the SLLN still holds for some non-homogeneous, not identically distributed Markov chains (see [19, p. 129–135]). Finally, Markov chains on countable sets satisfy the SLLNE if they are irreducible, positive recurrent, and homogeneous (see e.g. [24, Theorem 1.10.2]).

2.2. Loss functions, risks, and consistency

In this section we recall some basic notions for loss functions and their associated risks. We then introduce consistency notions for learning algorithms for stochastic processes satisfying a law of large numbers.

In the following X is always a measurable space if not mentioned otherwise and $Y \subset \mathbb{R}$ is always a closed subset. Moreover, metric spaces are always equipped with the Borel σ-algebra, and products of measurable spaces are always equipped with the corresponding product σ-algebra. Finally, $L_p(\mu)$ stands for the standard space of p-integrable functions with respect to the measure μ on X.

Definition 2.10. A function $L : X \times Y \times \mathbb{R} \to [0, \infty]$ is called a *loss function* if it is measurable. In this case we say that L convex (or continuous) if $L(x, y, \,.\,) : \mathbb{R} \to [0, \infty]$ is convex (or continuous) for all $x \in X, y \in Y$. Moreover, for a probability measure P on $X \times Y$ and an $f \in \mathcal{L}_0(X)$ the *L-risk* of f is defined by

$$\mathcal{R}_{L,P}(f) := \int_{X\times Y} L(x, y, f(x))\, \mathrm{d}P(x, y) = \int_X \int_Y L(x, y, f(x))\, \mathrm{d}P(y|x)\, \mathrm{d}P_X(x).$$

Finally, the *Bayes L-risk* is $\mathcal{R}^*_{L,P} := \inf\{\mathcal{R}_{L,P}(f) : f \in \mathcal{L}_0(X)\}$.

Note that the integral defining the L-risk always exists since L is non-negative and measurable. In addition it is obvious that the risk of a convex loss is convex on $\mathcal{L}_0(X)$. However, in general the risk of a continuous loss is not continuous. In order to ensure this continuity and several other, more sophisticated properties we need the following definition:

Definition 2.11. A loss function $L : X \times Y \times \mathbb{R} \to [0, \infty]$ is called a *Nemitski loss function* if there exist a measurable function $b : X \times Y \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ with

$$L(x, y, t) \le b(x, y) + h(|t|), \quad (x, y, t) \in X \times Y \times \mathbb{R}. \tag{9}$$

Furthermore, we say that L is a *Nemitski loss of order* $p \in (0, \infty)$, if there exists a constant $c > 0$ with $h(t) = c\, t^p$ for all $t \ge 0$. Finally, if P is a distribution on $X \times Y$ with $b \in L_1(P)$ we say that L is a *P-integrable Nemitski loss.*

Note that P-integrable Nemitski loss functions L satisfy $\mathcal{R}_{L,P}(f) < \infty$ for all $f \in L_\infty(P_X)$, and consequently we also have $\mathcal{R}_{L,P}(0) < \infty$ and $\mathcal{R}^*_{L,P} < \infty$.

For our further investigations we also need the following additional properties which are satisfied by basically all commonly used loss functions:

Definition 2.12. Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function. We say that L is:

(i) *locally bounded* if for all bounded $A \subset \mathbb{R}$ the restriction $L_{|X \times Y \times A}$ of L is a bounded function.
(ii) *locally Lipschitz continuous* if for all $a > 0$ we have

$$|L|_{a,1} := \sup_{\substack{t,t' \in [-a,a] \\ t \ne t'}} \sup_{\substack{x \in X \\ y \in Y}} \frac{|L(x, y, t) - L(x, y, t')|}{|t - t'|} < \infty. \tag{10}$$

(iii) *Lipschitz continuous* if we have $|L|_1 := \sup_{a>0} |L|_{a,1} < \infty$.

Note that if $Y \subset \mathbb{R}$ is a *finite* subset and $L : Y \times \mathbb{R} \to [0, \infty)$ is a convex loss function then L is a locally Lipschitz continuous loss function. Moreover, a locally Lipschitz continuous loss function L is a Nemitski loss since (10) yields

$$L(x, y, t) \le L(x, y, 0) + |L|_{|t|,1}|t|, \quad (x, y, t) \in X \times Y \times \mathbb{R}. \tag{11}$$

In particular, a locally Lipschitz continuous loss L is a P-integrable Nemitski loss if and only if $\mathcal{R}_{L,P}(0) < \infty$. Moreover, if L is Lipschitz continuous then L is a Nemitski loss of order 1.

The following examples recall that (locally) Lipschitz continuous losses are often used in learning algorithms for classification and regression problems:

Example 2.13. A loss $L : Y \times \mathbb{R} \to [0, \infty)$ of the form $L(y, t) = \varphi(yt)$ for a suitable function $\varphi : \mathbb{R} \to [0, \infty)$ and all $y \in Y := \{-1, 1\}$ and $t \in \mathbb{R}$, is called *margin-based*. Recall that margin-based losses such as the (squared) hinge loss, the AdaBoost loss, the logistic loss and the least squares loss are used in many classification algorithms. Obviously, L is convex, continuous, or (locally) Lipschitz continuous if and only if φ is. In addition, convexity of L implies local Lipschitz continuity of L. Moreover, L is always a P-integrable Nemitski loss since we have

$$L(y, t) \le \max\{\varphi(-t), \varphi(t)\} \tag{12}$$

for all $y \in Y$ and all $t \in \mathbb{R}$. In particular, this estimate shows that every convex margin-based loss is locally bounded. Moreover, from (12) we can easily derive a characterization for L being a P-integrable Nemitski loss of order p.

Example 2.14. A loss $L : Y \times \mathbb{R} \to [0, \infty)$ of the form $L(y, t) = \psi(y - t)$ for a suitable function $\psi : \mathbb{R} \to [0, \infty)$ and all $y \in Y := \mathbb{R}$ and $t \in \mathbb{R}$, is called *distance-based*. Distance-based losses such as the least squares loss, Huber's insensitive loss, the logistic loss, or the ϵ-insensitive loss are usually used for regression. Moreover, these examples illustrate that in general distance-based loss functions are neither locally bounded nor locally Lipschitz continuous. On the other hand, it is easy to see that L is convex, continuous, or Lipschitz continuous if and only if ψ is. Let us assume that L is a convex loss, i.e. ψ is convex. Then ψ is locally Lipschitz continuous and hence $V(r) := |\psi_{|[-r,r]}|_1$, where $|\psi_{|[-r,r]}|_1$ denotes the Lipschitz constant of the function $\psi_{|[-r,r]} : [-r, r] \to [0, \infty)$ is defined for all $r \ge 0$. Moreover, [4, Lemma 4] shows

$$V(r) \le \frac{2}{r} \|\psi_{|[-2r,2r]}\|_\infty \le 4V(2r), \quad r > 0. \tag{13}$$

Let us say that L is of *upper growth* $p \in [1, \infty)$ if there is a $c > 0$ with

$$\psi(r) \le c\,(|r|^p + 1), \quad r \in \mathbb{R}.$$

Analogously, L is said to be of *lower growth* $p \in [1, \infty)$ if there is a $c > 0$ with

$$\psi(r) \ge c\,(|r|^p - 1), \quad r \in \mathbb{R}.$$

Recall that most of the commonly used distance-based loss functions including the above examples are of the same upper and lower growth type. It is obvious that L is of upper growth type 1 if it is Lipschitz continuous, and if L is convex the converse implication also holds. Moreover, non-trivial convex L are always of lower growth type 1. In addition, a distance-based loss function of upper growth type $p \in [1, \infty)$ is a Nemitski loss of order p, and if the distribution P satisfies the moment condition

$$|P|_p := \left(\mathbb{E}_{(x,y)\sim P}|y|^p\right)^{1/p} := \left(\int_{X\times\mathbb{R}} |y|^p \, \mathrm{d}P(x,y)\right)^{1/p} < \infty \tag{14}$$

it is also P-integrable.

If our observations are realizations of a sequence $\mathcal{Z}$ of random variables $(X_i, Y_i) : \Omega \to X \times Y$ satisfying a law of large numbers then the following lemma proved in Section 4 shows that the risk with respect to the asymptotic mean distribution P actually describes the average future loss.

Lemma 2.15. *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, X be a measurable space, $Y \subset \mathbb{R}$ be a closed subset, and $\mathcal{Z} := ((X_i, Y_i))_{i\geq 1}$ be a $X \times Y$-valued stochastic process on Ω satisfying the WLLNE. Furthermore, let P be the asymptotic mean of $(\mathcal{Z}, \mu)$ and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function. If L is locally bounded then for all $f \in \mathcal{L}_\infty(X)$ and all $n_0 \geq 0$ we have*

$$\mathcal{R}_{L,P}(f) = \lim_{n\to\infty} \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} L\left(X_i, Y_i, f(X_i)\right), \tag{15}$$

where the limit is with respect to the convergence in probability μ. Moreover, if $\mathcal{Z}$ actually satisfies the SLLNE then (15) holds μ-almost surely. Finally, the same conclusions hold if L is a P-integrable Nemitski loss and $\mathcal{Z}$ satisfies the WLLN or SLLN.

With the help of the above lemma we can now introduce some concepts describing the asymptotic learning ability of learning algorithms. To this end recall that a method $\mathcal{L}$ that provides to every *training set* $T := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ a (measurable) function $f_T : X \to \mathbb{R}$ is called a *learning method*. The following definition introduces an asymptotic way to describe whether a learning method can learn from samples:

Definition 2.16. Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, X be a measurable space, $Y \subset \mathbb{R}$ be a closed subset, and $\mathcal{Z} := ((X_i, Y_i))_{i\geq 1}$ be a $X \times Y$-valued stochastic process on Ω satisfying the WLLNE. Furthermore, let P be the asymptotic mean of $(\mathcal{Z}, \mu)$ and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function. We say that a learning method $\mathcal{L}$ is *L-consistent for $\mathcal{Z}$* if

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_{T_n}) = \mathcal{R}^*_{L,P} \tag{16}$$

holds in probability μ, where $T_n := ((X_1, Y_1), \ldots, (X_n, Y_n))$ and $\mathcal{R}^*_{L,P}$ is the Bayes risk defined in Definition 2.10. Moreover, we say that $\mathcal{L}$ is *strongly L-consistent for $\mathcal{Z}$* if (16) holds μ-almost surely.

2.3. Consistency of SVMs

In this subsection we present some results showing that support vector machines (SVMs) can learn whenever the data-generating process satisfies a law of large numbers.

Let us begin by recalling the definition of SVMs. To this end let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function and H be a reproducing kernel Hilbert space (RKHS) over X (see e.g. [25]). Then for all $\lambda > 0$ and all observations $T := ((x_1, y_1), \ldots, (x_n, y_n)) \in X \times Y$ there exists exactly one element $f_{T,\lambda} \in H$ such that

$$f_{T,\lambda} \in \arg\min_{f\in H} \lambda\|f\|_H^2 + \frac{1}{n}\sum_{i=1}^{n} L\left(x_i, y_i, f(x_i)\right). \tag{17}$$

Given a null-sequence (λ_n) of strictly positive real numbers we call the learning method that provides to every training set $T \in (X \times Y)^n$ the decision function f_{T,λ_n} an *(λ_n)-SVM* based on H and L. For more information on SVMs we refer the reader to [26,27].

Moreover, given a distribution P on $X \times Y$ we say that the RKHS H is *(L, P)-rich* if we have

$$\mathcal{R}^*_{L,P,H} := \inf_{f\in H} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P},$$

i.e. if the Bayes risk can be approximated by functions from H. Note that the condition $\mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P}$ is satisfied (see [28]) whenever, the kernel of H is universal in the sense of [29], i.e. X is a compact metric space and H is dense in the space $C(X)$ of continuous functions. Less restrictive assumptions on H and X have been recently found in [28]. In particular, it was shown in [28] that the RKHSs H_σ, $\sigma > 0$, of the Gaussian RBF kernels

$$k_\sigma(x, x') := \exp\left(-\sigma^2\|x - x'\|_2^2\right), \quad x, x' \in \mathbb{R}^d$$

are (L, P)-rich for all distributions P on $\mathbb{R}^d \times Y$ and all continuous, P-integrable Nemitski losses L of order $p \in [1, \infty)$. Finally, one can also find some necessary and sufficient conditions for (L, P)-richness on countable spaces X in [28].

In order to present our first main result let us recall that a Polish space is a separable topological space whose topology can be described by a complete metric. It is well known that e.g. closed and open subset of $\mathbb{R}^d$ and compact metric spaces are Polish.

Now our first theorem essentially shows that there exists a consistent SVM for every process that takes values in a Polish space and that satisfies a law of large numbers for events.

Theorem 2.17. *Let X be a Polish space, $Y \subset \mathbb{R}$ be a closed subset and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, locally Lipschitz continuous, and locally bounded loss function. Moreover, let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{Z} := ((X_i, Y_i))_{i\geq 1}$ be an $X \times Y$-valued stochastic process on Ω satisfying the WLLNE, and P be the asymptotic mean of $(\mathcal{Z}, \mu)$. Finally, let H be an (L, P)-rich RKHS over X with bounded and continuous kernel. Then there exists a null-sequence (λ_n) of strictly positive real numbers such that the (λ_n)-SVM based on H and L is L-consistent for $\mathcal{Z}$.*

In addition, if $\mathcal{Z}$ satisfies the SLLNE then (λ_n) can be chosen such that the (λ_n)-SVM is strongly L-consistent for $\mathcal{Z}$.

We have seen in Example 2.14 that distance-based loss functions are in general are locally bounded. Nonetheless the following theorem establishes consistency for such losses.

Theorem 2.18. *Let X be a Polish space, $Y \subset \mathbb{R}$ be closed and $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss function of upper growth-type $p \in [1, \infty)$. Moreover, let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{Z} := ((X_i, Y_i))_{i\geq 1}$ be an $X \times Y$-valued stochastic process on Ω satisfying the WLLN, and P be the asymptotic mean of $(\mathcal{Z}, \mu)$. We assume that $|P|_p < \infty$. Finally, let H be the (L, P)-rich RKHS of a bounded and continuous kernel on X. Then there exists a null-sequence (λ_n) of strictly positive real numbers such that the (λ_n)-SVM based on H and L is L-consistent for $\mathcal{Z}$.*

In addition, if $\mathcal{Z}$ satisfies the SLLN then (λ_n) can be chosen such that the (λ_n)-SVM is strongly L-consistent for $\mathcal{Z}$.

The techniques used in the proofs of Theorems 2.17 and 2.18 are based on a (hidden) skeleton argument in the proof of Lemma 4.4. A more general though standard skeleton argument can be used to derive results similar to Theorems 2.17 and 2.18 for other empirical risk minimization methods using hypothesis sets with reasonably controllable complexity. Due to space constraints we omit the details.

Let us now assume for a moment that X is a subset of $\mathbb{R}^d$, L is a loss function in the sense of either Theorem 2.17 or 2.18, and H is the RKHS of a Gaussian RBF kernel. Then the above theorems together with the richness results from [28] show that for all data-generating processes $\mathcal{Z}$ satisfying a law of large numbers there exist suitable regularization sequences (λ_n) that allows us to build a consistent SVM. However, the sequences of Theorem 2.17 or 2.18 depend on $\mathcal{Z}$, and consequently, it would be desirable to have either a universal sequence (λ_n), i.e. a sequence that guarantees consistency for all $\mathcal{Z}$, or a *consistent* method that finds suitable values for λ from the observations. Unfortunately, the following theorem due to Nobel, [10], together with Birkhoff's ergodic theorem shows that neither of these alternatives is possible[2]:

Theorem 2.19. *There is no learning method which is L_{lsquares}-consistent for all stationary ergodic processes (X_i, Y_i) with values in $[0, 1] \times [0, 1]$, where L_{lsquares} denotes the usual least square loss $L_{\text{lsquares}}(y, t) := (y - t)^2$, $y, t \in \mathbb{R}$. Moreover, there is no learning method which is L_{class}-consistent for all stationary ergodic processes (X_i, Y_i) with values in $[0, 1] \times \{-1, 1\}$, where L_{class} denotes the classification loss $L_{\text{class}}(y, t) := \mathbf{1}_{(-\infty,0]}(y \operatorname{sign} t)$, $y = \pm 1$, $t \in \mathbb{R}$.*

Roughly speaking the impossibility of finding a universal sequence (λ_n) is related to the fact that there is no *uniform* convergence speed in the LLNs for general processes. More precisely, if $\mathcal{Z} := ((X_i, Y_i))_{i\geq 1}$ is a stochastic process which satisfies a law of large numbers then for all $\varepsilon > 0$, $n \geq 1$, and all suitable functions $f : X \times Y \to \mathbb{R}$ there exists a $\delta(\varepsilon, f, n) > 0$ with

$$\mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^{n} f \circ (X_i, Y_i)(\omega) - \mathbb{E}_P f\right| > \varepsilon\right\}\right) \leq \delta(\varepsilon, f, n) \tag{18}$$

and $\lim_{n\to\infty} \delta(\varepsilon, f, n) = 0$. Now, the proofs of Theorems 2.17 and 2.18 (essentially) show that we can determine a sequence (λ_n) whenever we know such $\delta(\varepsilon, f, n)$ for all $\varepsilon > 0$, $n \geq 1$, and a suitably large class of functions f. However, since there exists no universal sequence (λ_n) by Theorem 2.19 we consequently see that there exists no values $\delta(\varepsilon, f, n)$ such that (18) holds for *all* (stationary) processes satisfying a law of large numbers.

This discussion shows that in order to build consistent SVMs for interesting classes of processes one has to find *quantitative* versions of laws of large numbers. In the following section we will present a simple yet powerful method for establishing such versions for mixing processes.

[2] Recall that binary classification is the "easiest" non-parametric learning problem in the sense that negative results for this learning problem can typically be translated into negative results for almost all learning problems defined by loss functions (cf. p. 118f in [30] for some examples in this direction and the proof of the below theorem in [10] for the least squares loss).

3. Consistency for mixing processes

In this section we derive consistency results for SVMs under the assumption that the data-generating process satisfies certain mixing conditions. These mixing conditions generally quantify how much a process fails to be independent. In the first subsection we recall some commonly used mixing conditions. In the second subsection we then present our consistency results and compare them with known consistency results for other learning algorithms.

3.1. Mixing coefficients for processes

In this subsection we recall some standard mixing coefficients and their basic properties (see e.g. [31,17] for thorough treatment). To this end let Ω be a set, $\mathcal{A}$ and $\mathcal{B}$ be two σ-algebras on Ω, and μ be a probability measure on $\sigma(\mathcal{A} \cup \mathcal{B})$. Furthermore, let H be a Hilbert space and $\mathcal{L}_p(\mathcal{A}, \mu, H)$ be the space of all $\mathcal{A}$-measurable H-valued functions that are p-integrable with respect to μ. Using the convention $\frac{0}{0} := 0$ we define the following mixing coefficients for the pair $(\mathcal{A}, \mathcal{B})$:

$$\alpha(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A\in\mathcal{A}\\ B\in\mathcal{B}}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

$$\varphi(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{A\in\mathcal{A}\\ B\in\mathcal{B}}} \left|\frac{\mu(A \cap B) - \mu(A)\mu(B)}{\mu(A)}\right|$$

$$\varphi_{\mathrm{sym}}(\mathcal{A}, \mathcal{B}, \mu) := \sqrt{\varphi(\mathcal{A}, \mathcal{B}, \mu) \cdot \varphi(\mathcal{B}, \mathcal{A}, \mu)}$$

$$R_p^H(\mathcal{A}, \mathcal{B}, \mu) := \sup_{\substack{f\in\mathcal{L}_p(\mathcal{A},\mu,H)\\ g\in\mathcal{L}_p(\mathcal{B},\mu,H)}} \left|\frac{\mathbb{E}_\mu\langle f, g\rangle - \langle \mathbb{E}_\mu f, \mathbb{E}_\mu g\rangle}{\|f\|_p \, \|g\|_p}\right|, \quad p \in [2, \infty].$$

It is obvious from the definitions that all mixing coefficients equal 0 if $\mathcal{A}$ and $\mathcal{B}$ are independent, and besides φ they are also symmetric in $\mathcal{A}$ and $\mathcal{B}$. Moreover, we have $2\alpha(\mathcal{A}, \mathcal{B}, \mu) \le \varphi(\mathcal{A}, \mathcal{B}, \mu)$ and $4\alpha(\mathcal{A}, \mathcal{B}, \mu) \le R_p^{\mathbb{R}}(\mathcal{A}, \mathcal{B}, \mu) \le 2\varphi_{\mathrm{sym}}(\mathcal{A}, \mathcal{B}, \mu)$ for all $p \in [2, \infty]$, see [31, Section 1] and the references therein. Furthermore, [32, Theorem 4.1] shows that for all $p \in [2, \infty]$ there exists a constant $c_p > 0$ such that for all Hilbert spaces H we have

$$R_p^{\mathbb{R}}(\mathcal{A}, \mathcal{B}, \mu) \le R_p^H(\mathcal{A}, \mathcal{B}, \mu) \le c_p \, R_p^{\mathbb{R}}(\mathcal{A}, \mathcal{B}, \mu). \tag{19}$$

Note that for $p = 2$ we actually have $c_p = 1$ and for $p = \infty$ we may choose the famous Grothendieck constant (see the proof of Lemma 2.2 in [33]). Moreover, it is obvious from the definition that $R_p^H(\mathcal{A}, \mathcal{B}, \mu)$ is decreasing in p, i.e. $R_p^H(\mathcal{A}, \mathcal{B}, \mu) \le R_q^H(\mathcal{A}, \mathcal{B}, \mu)$ for $q \le p$. Finally, Theorem 4.13 in [34] gives the highly non-trivial relation

$$R_p^{\mathbb{R}}(\mathcal{A}, \mathcal{B}, \mu) \le 2\pi \, \alpha^{1-\frac{2}{p}}(\mathcal{A}, \mathcal{B}, \mu) \, \varphi_{\mathrm{sym}}^{\frac{2}{p}}(\mathcal{A}, \mathcal{B}, \mu), \quad p \in [2, \infty]. \tag{20}$$

Let us now consider mixing coefficients and corresponding mixing notions for stochastic processes:

Definition 3.1. Let Z be a Z-valued stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$ and let ξ be one of the above mixing coefficients. For $i, j \ge 1$ we define the *ξ-bi-mixing coefficient of Z* by

$$\xi(Z, \mu, i, j) := \xi\left(\sigma(Z_i), \sigma(Z_j), \mu\right).$$

Furthermore, for $n \ge 1$ the *ξ-mixing* and *$\bar{\xi}$-mixing* coefficients of Z are defined by

$$\xi(Z, \mu, n) := \sup_{i\ge 1} \xi(Z, \mu, i, i+n)$$

$$\bar{\xi}(Z, \mu, n) := \sup_{i\ge 1} \xi\left(\sigma(Z_1, \ldots, Z_i), \sigma(Z_{i+n}, Z_{i+1+n}, \ldots), \mu\right).$$

It is immediately clear that $\xi(Z, \mu, n) \le \bar{\xi}(Z, \mu, n)$. This trivial observation is interesting since the literature typically deals with $\bar{\xi}(Z, \mu, n)$, whereas the consistency results which we will present in the following subsection only require bounds on $\xi(Z, \mu, n)$ or $\xi(Z, \mu, i, j)$. Finally, recall that for stationary, homogeneous Markov chains Z we actually have $\xi(Z, \mu, n) = \bar{\xi}(Z, \mu, n)$ if $\xi \ne \varphi_{\mathrm{sym}}$.

In the following we say that the process Z is *ξ-mixing* if $\lim_{n\to\infty} \xi(Z, \mu, n) = 0$. Moreover, Z is called *weakly ξ-mixing* if $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \xi(Z, \mu, k) = 0$. In addition, we define analogous mixing notions for $\bar{\xi}$. Finally, Z is said to be *weakly ξ-bi-mixing* if

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{i-1} \xi(Z, \mu, i, j) = 0. \tag{21}$$

Obviously, every ξ-mixing process is weakly ξ-mixing, and since a simple induction over $n \in \mathbb{N}$ shows

$$\sum_{i=1}^{n}\sum_{j=1}^{i-1}\xi(\mathcal{Z},\mu,i,j)=\sum_{k=1}^{n-1}\sum_{m=1}^{n-k}\xi(\mathcal{Z},\mu,m+k,m),\quad n\geq 1,$$

we also see that every weakly ξ-mixing process is weakly ξ-bi-mixing. Moreover, if the process $\mathcal{Z}$ is stationary in the wide sense then an elementary proof shows that $\xi(\mathcal{Z},\mu,i,j)=\xi(\mathcal{Z},\mu,i+k,j+k)$. Since this implies $\xi(\mathcal{Z},\mu,i,j)=\xi(\mathcal{Z},\mu,i-j+1)$ for $i\geq j\geq 1$ we then find

$$\sum_{i=1}^{n}\sum_{j=1}^{i-1}\xi(\mathcal{Z},\mu,i,j)=\sum_{k=1}^{n-1}\sum_{m=1}^{n-k}\xi(\mathcal{Z},\mu,m+k,m)=\sum_{k=1}^{n-1}(n-k)\,\xi(\mathcal{Z},\mu,k+1). \tag{22}$$

Consequently, every *stationary* weakly ξ-bi-mixing process is actually weakly ξ-mixing.

Some information on mixing conditions for stationary processes and their relation to mixing in the ergodic sense can be found in e.g. [31]. Examples of (exponentially or polynomially) $\bar{\xi}$-mixing processes including certain Markov, ARMA, MA(∞), and GARCH processes can be found in [35, Sect. 2.6.1] and [31,34,36]. Moreover, mixing properties of Gaussian processes are considered in [34, Chapter 9]. Finally, [37, Theorem 26.5] together with [34, Proposition 3.18] shows that in general the $\bar{\xi}$-mixing rates can be arbitrarily slow. A brief survey of these and other results together with various references is given in [31], and a thorough and recent treatment can be found in [34,36,37].

Let us finally discuss some laws of large numbers for mixing processes. We begin with the following simple result proved in Section 4 which shows that asymptotically mean stationary, weakly bi-mixing processes satisfy the WLLNE:

Proposition 3.2. *Let $\mathcal{Z}$ be a Z-valued weakly α-bi-mixing stochastic process. Then $\mathcal{Z}$ is AMS if and only if it satisfies the WLLNE.*

Using [19, Theorm 8.2.1] it is easy to see that for $\bar{\alpha}$-mixing processes being AMS is actually equivalent to the SLLNE. Finally, [38, Cor. 8.2.2] shows that *identically distributed* processes $\mathcal{Z}$ with

$$\sum_{n=1}^{\infty}\sqrt{\bar{\varphi}(\mathcal{Z},\mu,2^n)}<\infty \tag{23}$$

satisfy the SLLN. Obviously, (23) is satisfied whenever there are constants $c>0$ and $\alpha>2$ such that $\bar{\varphi}(\mathcal{Z},\mu,n)\leq c\,(\ln n)^{-\alpha}$ for all $n\geq 2$.

3.2. Consistency of SVMs for mixing processes

In this subsection we establish consistency results for data-generating processes with known upper bounds on the weakly α-bi-mixing rate. Unlike in the case of general processes satisfying a law of large numbers these new consistency results give explicit conditions on the regularization sequences guaranteeing consistency.

In order to formulate these results we have to introduce a new quantity. To this end let k be a bounded kernel over some set X. Then the supremum norm of k is defined by

$$\|k\|_\infty:=\sup_{x\in X}\sqrt{k(x,x)}.$$

Note that for the Gaussian kernels k_σ we have $\|k_\sigma\|_\infty=1$.

Now we can present our first consistency result which deals with locally Lipschitz-continuous loss functions:

Theorem 3.3. *Let X be a separable metric space, $Y\subset\mathbb{R}$ be a closed subset and $L: X\times Y\times\mathbb{R}\to[0,\infty)$ be a convex, locally Lipschitz continuous loss function with $\|L(.,.,0)\|_\infty\leq c$. Moreover, let $(\Omega,\mathcal{A},\mu)$ be a probability space, $\mathcal{Z}:=((X_i,Y_i))_{i\geq 1}$ be an $X\times Y$-valued, AMS stochastic process on Ω, and P be the asymptotic mean of $(\mathcal{Z},\mu)$. In addition, let H be an (L,P)-rich RKHS over X with bounded continuous kernel k. We write*

$$B_\lambda:=\|k\|_\infty\left(\frac{c}{\lambda}\right)^{1/2},\quad \lambda>0.$$

Finally, assume that there are constants $C\in(0,\infty)$ and $\alpha\in(0,1]$ such that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\mu f\circ Z_i-\mathbb{E}_P f\right|\leq C\|f\|_\infty n^{-\alpha} \tag{24}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\alpha(\mathcal{Z},\mu,i,j)\leq Cn^{-\alpha} \tag{25}$$

for all $f\in\mathcal{L}_\infty(Z)$ and all $n\geq 1$. Then for all null-sequence (λ_n) of strictly positive real numbers satisfying

$$\frac{|L|^4_{B_{\lambda_n},1}}{\lambda_n^2 n^\alpha}\to 0 \tag{26}$$

the corresponding (λ_n)-SVM based on H and L is L-consistent for $\mathcal{Z}$.

The above result is of particular interest for binary classification problems. Indeed, recall that the standard SVM for classification uses the hinge loss defined by

$$L(y, t) := \max\{0, 1 - yt\}, \quad y \in Y := \{-1, 1\},\ t \in \mathbb{R}.$$

Obviously, this loss function is convex and Lipschitz continuous with $|L|_1 = 1$ and $L(y, 0) = 1$ for $y \in Y$. For $X := \mathbb{R}^d$ and H_σ being the RKHS of a Gaussian RBF kernel with fixed width σ we consequently obtain L-consistency for the corresponding (λ_n)-SVM whenever $\lambda_n \to 0$ and $\lambda_n^2 n^\alpha \to \infty$, where α is the exponent satisfying (24) and (25). Since L-consistency implies binary classification consistency (see e.g. [3,39]) we hence see that the above SVM is classification consistent. Note that this consistency generalizes earlier consistency results of [1–3] with respect to both the compactness assumption on X and the i.i.d. assumption on the data-generating process. Finally, in the case of $\alpha = 1$ the SVMs using the hinge loss L and an (L, P)-rich RKHS is consistent if $\lambda_n \to 0$ and $n\lambda_n^2 \to \infty$. Since this is exactly the condition ensuring consistency in the i.i.d. case we see that such an SVM is quite robust against violations of the i.i.d. assumption.

If quantitative approximation properties of H in terms of convergence rates for $\mathcal{R}_{L,P}(f_{P,\lambda}) \to \mathcal{R}^*_{L,P}$ are known, the proof Theorem 3.3 also provides learning rates. However, we conjecture that these rates are usually overly conservative in terms of the estimation error, i.e. the statistical part of the analysis, since for the latter we only employ Markov's inequality in a very straightforward fashion. Nonetheless, sharper learning rates seem to be possible for e.g. exponentially β-mixing processes such as certain Markov chains. However, our experience from the analysis for i.i.d. processes suggests that quite involved techniques are needed to obtain *sharp learning rates* (and not only sharp rates for the estimation error). For example, [40] shows that in order to correctly combine the approximation error with the estimation error a localization argument with respect to the regularization parameter is needed. Moreover, it is well known that for i.i.d. processes the so-called variance bounds can drastically improve the learning rates, and such variance bounds typically lead to another localization argument that uses Talagrand's inequality. For these reasons we feel that any serious consideration of learning rates is out of the scope of the paper. Instead we would like to compare our consistency result with the consistency result for regularized boosting algorithms derived in [41]. To this end we first observe that for (in the wide sense) stationary processes (24) is automatically satisfied and (25) is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} \alpha(\mathcal{Z}, \mu, i) \le Cn^{-\alpha}, \quad n \ge 1,$$

by (22). Obviously, the latter is satisfied if $\mathcal{Z}$ is *algebraically $\bar{\alpha}$-mixing with exponent α*, i.e. if it satisfies $\bar{\alpha}(\mathcal{Z}, \mu, n) \le Cn^{-\alpha}$ for all $n \ge 1$. Consequently, Theorem 3.3 implies consistency results for stationary, algebraically $\bar{\alpha}$-mixing processes for which we have a bound on the mixing rate. Compared to this [41] only establishes a consistency result for stationary, algebraically $\bar{\beta}$-mixing processes for which we have a bound on the mixing rate. Since in general $\bar{\alpha}$-mixing is strictly weaker assumption than $\bar{\beta}$-mixing we see that Theorem 3.3 substantially weakens the assumptions of [41]. Finally, note that our restriction to polynomial rates in (24) and (25) is by no means necessary. For example, if we replace $n^{-\alpha}$ by $(\log n)^{-\alpha}$ in (24) and (25) then the corresponding condition on (λ_n) for the SVM using the hinge loss becomes $\lambda_n^2(\log n)^\alpha \to \infty$. In particular, note that such an SVM is consistent for *all* stationary, algebraically α-mixing processes![3] In this direction it is interesting to recall that in [12] consistency was established for kernel estimators and algebraically α-mixing, not necessarily stationary processes. To the best of our knowledge this is the consistency result that is closest in its assumptions on $\mathcal{Z}$ to Theorem 3.3.

The proof of Theorem 3.3 is based on a stability argument together with a simple Markov-type concentration inequality for Hilbert-space-valued random variables. In principle, one could also employ exponential type inequalities for sums of $\mathbb{R}$-valued random variables in the sense of e.g. [17, Chapter 1.4] together with a skeleton argument based on e.g. covering numbers. However, some preliminary considerations we made in this direction suggest that at least for a straightforward approach the resulting conditions on (λ_n) are substantially stronger. Consequently, we do not discuss such an approach in further detail.

The next theorem establishes a result similar to Theorem 3.3 for distance-based loss functions of some growth type p:

Theorem 3.4. *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a convex distance-based loss function of upper growth type $p \in [1, 2]$. Furthermore, let X be a separable metric space and H be an (L, P)-rich RKHS over X with bounded continuous kernel k. Moreover, let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{Z} := ((X_i, Y_i))_{i\ge1}$ be an $X \times \mathbb{R}$-valued, AMS stochastic process on Ω, and P be the asymptotic mean of $(\mathcal{Z}, \mu)$. Assume that we have*

$$\sup_{i\ge1} |\mu_{(X_i,Y_i)}|_q < \infty \tag{27}$$

for some $q \in [p, \infty]$, where $|.|_q$ is the moment defined by (14). Furthermore assume that there are constants $C > 0$ and $\alpha, \beta \in (0, 1]$ such that

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\mu f \circ Z_i - \mathbb{E}_P f\right| \le C\|f\|_{L_1(P)}\, n^{-\alpha} \tag{28}$$

[3] However, for such (λ_n) the SVM typically deals too conservatively with the stochastic part of the learning process, so that the approximation behaviour is poor. As a consequence this result does not seem to have any practical relevance.

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\alpha^{1-\frac{2p-2}{q}}(\mathcal{Z},\mu,i,j)\varphi_{\mathrm{sym}}^{\frac{2p-2}{q}}(\mathcal{Z},\mu,i,j)\leq Cn^{-\beta} \tag{29}$$

for all $f\in L_1(P)\cap\bigcap_{i=1}^{\infty}L_1(\mu_{(X_i,Y_i)})$. *Then for all null-sequences* (λ_n) *of strictly positive real numbers satisfying the conditions*

$$\lambda_n^p n^{2\alpha}\to\infty \tag{30}$$

$$\lambda_n^{2p} n^{\beta}\to\infty \tag{31}$$

the corresponding (λ_n)*-SVM based on* H *and* L *is* L*-consistent for* $\mathcal{Z}$.

Since distance-based loss functions are typically used for regression problems we see that the above theorem is mainly interesting for these learning scenarios. For Lipschitz continuous losses such as the absolute distance loss $L(y,t):=|y-t|$, the ϵ-insensitive loss $L(y,t):=\max\{0,|y-t|-\epsilon\}$, the logistic loss or Huber's robust loss we obviously have $p=1$ and hence (29) reduces to (25). Moreover, for Lipschitz continuous losses we can choose $q=1$ in (27). Consequently, it is easy to see that all remarks made for the classification SVM using the hinge loss, remain true for regression SVMs using one of the above losses.

In contrast to this an SVM that uses the standard least squares loss requires $p=2$ in the above theorem. For processes with uniformly bounded noise, i.e. $q=\infty$, we again see that (29) reduces to (25). Moreover, for $q\in(2,\infty)$ we have

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\alpha^{1-\frac{2}{q}}(\mathcal{Z},\mu,i,j)\varphi_{\mathrm{sym}}^{\frac{2}{q}}(\mathcal{Z},\mu,i,j)\leq\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\alpha(\mathcal{Z},\mu,i,j)\right)^{1-\frac{2}{q}}$$

so that (25) implies (29) for $\beta:=\alpha(1-2/q)$. However, for $q=2$ we have $1-\frac{2p-2}{q}=0$, and consequently we only obtain consistency results for weakly φ_{sym}-bi-mixing processes.

Theorem 3.4 generalizes the only known consistency result (see [4]) for regression SVMs dealing with unbounded noise with respect to both the compactness assumption on X and the i.i.d. assumption on the data-generating process. In particular, Theorem 3.4 shows that such SVMs are rather robust against violations of these assumptions, and consequently it gives a strong justification of using such SVMs in rather general situations.

Finally, we would like to mention that condition (27) can be replaced by a weaker assumption describing the average behaviour of the sequence $(|\mu_{(X_i,Y_i)}|_q)_{i\geq1}$. However, the resulting conditions on (λ_n) are more complicated and hence we omit the details.

4. Proofs

4.1. Proofs from Section 2.1

Proof of Lemma 2.3. Let $\mathcal{B}$ be the σ-algebra of Z. We write $P_n(B):=\frac{1}{n}\sum_{i=1}^{n}\mu(Z_i\in B)$ for $B\in\mathcal{B}$ and $n\geq1$. Then P_n is obviously a probability measure on $\mathcal{B}$ for all $n\geq1$. Now the theorem of Vitali–Hahn–Saks (see e.g. [42, p. 158–160]) ensures that $P(B):=\lim_{n\to\infty}P_n(B)$, $B\in\mathcal{B}$, defines a probability measure on $\mathcal{B}$. ■

Proof of Theorem 2.4. Recall that the convergence in probability μ can be described by the metric

$$d(f,g):=\int_{\Omega}\min\{1,|f-g|\}\,\mathrm{d}\mu,\quad f,g\in\mathcal{L}_0(\Omega).$$

Moreover, for measurable $B\subset Z$ let c_B be the constant satisfying (1). The WLLNE and the above metric then shows

$$\lim_{n\to\infty}\int_{\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_B\circ Z_i-c_B\right|\mathrm{d}\mu=0.$$

Since $\|.\|_{L_1(\mu)}$ is continuous on $L_1(\mu)$ we hence find

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\mu\mathbf{1}_B\circ Z_i=\lim_{n\to\infty}\int_{\Omega}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_B\circ Z_i\mathrm{d}\mu=\lim_{n\to\infty}\int_{\Omega}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_B\circ Z_i\right|\mathrm{d}\mu=\mathbb{E}_\mu|c_B|=c_B,$$

where the existence of the right limit implies the existence of the left limit. Consequently, $\mathcal{Z}$ is AMS and we have $P(B)=c_B$. Obviously, the latter together with (1) immediately gives (4). Finally, if $\mathcal{Z}$ satisfies the SLLNE then we obtain the almost sure convergence in (4) from (2). ■

Proof of Lemma 2.5. Let us begin by showing the assertion for the SLLNE. To this end we fix an $\varepsilon>0$. By the approximation lemma for bounded measurable functions there exists a step function $g:X\to\mathbb{R}$ with $\|f-g\|_\infty\leq\varepsilon$. Now, the linearity of the limit together with the SLLNE shows

$$\mathbb{E}_Pg=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}g\circ Z_i(\omega)$$

for μ-almost all $\omega \in \Omega$, and consequently, [43, Lemma 20.6] gives an $n_0 \geq 1$ such that

$$\mu\left(\sup_{n\geq n_0}\left|\frac{1}{n}\sum_{i=1}^{n} g\circ Z_i - \mathbb{E}_P g\right| \leq \varepsilon\right) \geq 1-\varepsilon. \tag{32}$$

Moreover, for $\omega \in \Omega$ the triangle inequality together with $\|f-g\|_\infty \leq \varepsilon$ yields

$$\sup_{n\geq n_0}\left|\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i(\omega) - \mathbb{E}_P f\right| \leq 2\varepsilon + \sup_{n\geq n_0}\left|\frac{1}{n}\sum_{i=1}^{n} g\circ Z_i(\omega) - \mathbb{E}_P g\right|,$$

and hence we obtain

$$\mu\left(\sup_{n\geq n_0}\left|\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i - \mathbb{E}_P f\right| \leq 3\varepsilon\right) \geq 1-\varepsilon.$$

This shows the μ-almost sure convergence in (5). Using that the functions $\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i$, $n \geq 1$, are uniformly bounded Lebesgue's theorem then yields

$$\mathbb{E}_P f = \int_\Omega \mathbb{E}_P f \,\mathrm{d}\mu = \int_\Omega \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f\circ Z_i \mathrm{d}\mu = \lim_{n\to\infty}\int_\Omega \frac{1}{n}\sum_{i=1}^{n} f\circ Z_i \mathrm{d}\mu = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\mu f\circ Z_i,$$

and hence we have found (6). Finally, if $\mathcal{Z}$ only satisfies the WLLNE then pulling the supremum out of μ in (32) and adjusting the rest of the proof accordingly shows (5) with convergence in probability μ. Moreover, in this case (6) can be shown analogously to the argument used in the proof Theorem 2.4. ■

4.2. Proofs from Section 2.2

Proof of Lemma 2.15. Let us first assume that L is locally bounded. It is then straightforward to check that it suffices to consider the case $n_0 = 0$. Now observe that the function $g(x,y) := L(x,y,f(x))$, $(x,y) \in X\times Y$, is a bounded, measurable function since f is assumed to be bounded, and L is locally bounded. Applying Lemma 2.5 to the function g then gives the assertion.

Let us now assume that L is a P-integrable Nemitski loss. Then there exist a $b \in L_1(P)$ and an increasing function $h : [0,\infty) \to [0,\infty)$ with

$$g(x,y) \leq b(x,y) + h\left(\|f\|_\infty\right), \quad (x,y) \in X\times Y.$$

This shows that $g \in L_1(P)$, and hence the assertion follows from Definition 2.6. ■

4.3. Proofs from Section 2.3

For the proof of Theorem 2.17 we need some preparations. Let us begin with the following result on the existence and uniqueness of infinite sample SVMs which is a slight extension of similar results established in [44,4]:

Theorem 4.1. *Let $L : X\times Y\times\mathbb{R}\to[0,\infty)$ be a convex loss function and P be a distribution on $X\times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be a RKHS of a bounded measurable kernel over X. Then for all $\lambda > 0$ there exists exactly one element $f_{P,\lambda} \in H$ such that*

$$\lambda\|f_{P,\lambda}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda}) = \inf_{f\in H}\lambda\|f\|_H^2 + \mathcal{R}_{L,P}(f). \tag{33}$$

Furthermore, we have $\|f_{P,\lambda}\|_H \leq \sqrt{\frac{\mathcal{R}_{L,P}(0)}{\lambda}}$.

The following two results describe the stability of the empirical SVM solutions. The first result was (essentially) shown in [44,4]:

Theorem 4.2. *Let X be a separable metric space, $L : X\times Y\times\mathbb{R}\to[0,\infty)$ be a convex, locally Lipschitz continuous loss function, and P be a distribution on $X\times Y$ with $\mathcal{R}_{L,P}(0) < \infty$. Furthermore, let H be the RKHS of a bounded, continuous kernel k over X with canonical feature map $\Phi : X\to H$. We define*

$$B_\lambda := \|k\|_\infty\left(\frac{\mathcal{R}_{L,P}(0)}{\lambda}\right)^{1/2}, \quad \lambda > 0.$$

Then for all $\lambda > 0$ there exists a bounded, measurable function $h_\lambda : X\times Y\to\mathbb{R}$ with

$$\|h_\lambda\|_\infty \leq |L|_{B_\lambda,1} \tag{34}$$

and

$$\|f_{P,\lambda} - f_{T,\lambda}\|_H \le \frac{1}{\lambda} \|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_T h_\lambda \Phi\|_H \tag{35}$$

for all training sets $T = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, *where* $\mathbb{E}_T$ *denotes the expectation operator with respect to the empirical measure associated to* T, *i.e.* $\mathbb{E}_T g := \frac{1}{n}\sum_{i=1}^n g(x_i, y_i)$.

Recall that convex distance-based loss functions are in general *not* locally Lipschitz continuous. Nevertheless SVM using these losses still enjoy stability as the following result shows:

Theorem 4.3. *Let* X *be a separable metric space,* $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ *be a convex, distance-based loss function of upper growth type* $p \ge 1$ *and* P *a distribution on* $X \times \mathbb{R}$ *with* $|P|_q < \infty$ *for some* $q \in [p, \infty]$. *Furthermore, let* H *be a RKHS of a bounded, continuous kernel over* X *with canonical feature map* $\Phi : X \to H$. *Then there exists a constant* $c_L > 0$ *depending only on* L *such that for all* $\lambda > 0$ *there exists a measurable function* $h_\lambda : X \times Y \to \mathbb{R}$ *with*

$$\|h_\lambda\|_{L_s(\bar{P})} \le 8^p c_L \left(1 + |\bar{P}|_q^{p-1} + \|f_{P,\lambda}\|_\infty^{p-1}\right) \tag{36}$$

$$\|f_{P,\lambda} - f_{T,\lambda}\|_H \le \frac{1}{\lambda} \|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_T h_\lambda \Phi\|_H \tag{37}$$

for $s := \frac{q}{p-1}$, *all distributions* $\bar{P}$ *on* $X \times \mathbb{R}$ *with* $|\bar{P}|_q < \infty$ *and all training sets* $T \in (X \times Y)^n$. *Finally, if* L *is also of lower growth type* p *then we additionally have*

$$\|h_\lambda\|_{L_s(P)} \le 16^p c_L \left(1 + |P|_q^{p-1}\right)\left(1 + \|f_{P,\lambda}\|_\infty^{\frac{q-p}{s}}\right). \tag{38}$$

Proof. By taking care in the constants in the proof of [4, Theorem 10] we obtain a measurable function $h_\lambda : X \times Y \to \mathbb{R}$ satisfying (37) and

$$|h_\lambda(x, y)| \le 4^p c_L \max\left\{1, |y - f_{P,\lambda}(x)|^{p-1}\right\}, \quad (x, y) \in X \times Y,$$

where c_L is a suitable constant depending only on the loss function L. For $q = \infty$ we then easily find the assertion, and hence let us assume that $q \in [p, \infty)$. In this case, the above inequality yields

$$|h_\lambda(x, y)|^s \le 4^{ps} c_L^s \max\left\{1, |y - f_{P,\lambda}(x)|^q\right\} \le 4^{ps} 2^{q-1} c_L^s \left(1 + |y|^q + |f_{P,\lambda}(x)|^q\right). \tag{39}$$

Since $\frac{q-1}{s} \le p$ and $s \ge 1$ we then obtain (36). Moreover, if ψ is the function satisfying $L(y, t) = \psi(y - t)$, $y, t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}_P|f_{P,\lambda}|^p &\le 2^{p-1}\int_{X\times Y} |y - f_{P,\lambda}(x)|^p + |y|^p \, \mathrm{d}P(x, y) \\
&\le 2^{p-1}\int_{X\times Y} c_L^{(1)}\, \psi\left(y - f_{P,\lambda}(x)\right) + 1 + |y|^p \, \mathrm{d}P(x, y) \\
&= 2^{p-1}\left(c_L^{(1)} \mathcal{R}_{L,P}(f_{P,\lambda}) + 1 + |P|_p^p\right) \\
&\le 2^{p-1}\left(c_L^{(1)} \mathcal{R}_{L,P}(0) + 1 + |P|_p^p\right) \\
&\le 2^{p-1}\left(c_L^{(2)}\left(1 + |P|_p^p\right) + 1 + |P|_p^p\right) \\
&\le 2^p c_L^{(3)}\left(1 + |P|_p^p\right),
\end{aligned}$$

where $c_L^{(1)}, c_L^{(2)} \ge 1$, and $c_L^{(3)} \ge 1$ are suitable constants depending only on the loss function L. Combining the estimate on $\mathbb{E}_P|f_{P,\lambda}|^p$ with (39) then gives

$$\begin{aligned}
\|h_\lambda\|_{L_s(P)} &\le 4^p\, 2^{\frac{q-1}{s}} c_L \left(1 + |P|_q^{p-1} + \|f_{P,\lambda}\|_\infty^{\frac{q-p}{s}} \left(\mathbb{E}_P|f_{P,\lambda}|^p\right)^{\frac{1}{s}}\right) \\
&\le 4^p\, 2^{\frac{q-1}{s}} c_L \left(1 + |P|_q^{p-1} + \|f_{P,\lambda}\|_\infty^{\frac{q-p}{s}} \left(2^p c_L^{(3)}(1 + |P|_p^p)\right)^{\frac{1}{s}}\right) \\
&\le 4^p\, 2^{\frac{p+q}{s}} \left(c_L^{(4)}\right)^{1+\frac{1}{s}} \left(1 + |P|_p^{\frac{p}{s}} + |P|_q^{p-1}\right)\left(1 + \|f_{P,\lambda}\|_\infty^{\frac{q-p}{s}}\right),
\end{aligned}$$

where $c_L^{(4)} \ge 1$ is another suitable constant depending only on the loss function L. Now note that we have $\frac{p+q}{s} = (\frac{p}{q} + 1)(p - 1) \le 2(p - 1)$ and $1 + \frac{1}{s} \le 2$. These estimates together with

$$|P|_p^{\frac{p}{s}} \le |P|_q^{\frac{p}{s}} = |P|_q^{\frac{p(p-1)}{q}} \le 1 + |P|_q^{p-1}$$

then yield (38). ■

The next lemma establishes Hilbert-space-valued laws of large numbers which are later used to bound the term $\|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_T h_\lambda \Phi\|_H$.

Lemma 4.4. *Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, Z be a Polish space, and $\mathcal{Z} := (Z_i)_{i\geq 1}$ be a Z-valued stochastic process on Ω. Assume that $\mathcal{Z}$ satisfies the WLLNE and let P be the asymptotic mean of $(\mathcal{Z}, \mu)$. Furthermore, let H be a Hilbert space, and $\Phi : Z \to H$ be a continuous and bounded map. Then for all $h \in L_\infty(P)$ we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (h\Phi)\circ Z_i = \mathbb{E}_P h\Phi, \tag{40}$$

where the convergence is in probability μ. Moreover, if $\mathcal{Z}$ actually satisfies the WLLN then (40) holds for all $f \in L_1(P)$. Finally, the convergence holds μ-almost surely for all $f \in L_\infty(P)$ or $f \in L_1(P)$ if $\mathcal{Z}$ satisfies the SLLNE or SLLN, respectively.

Proof. Let us first show (40) for $f \in L_1(P)$ when $\mathcal{Z}$ satisfies the SLLN. To this end we first make the additional assumption that there exists a compact subset $K \subset Z$ with $h(z) = 0$ for all $z \notin K$. Now recall that Φ is continuous and hence $\Phi(K) \subset H$ is compact. Moreover, recall that H as a Hilbert space has the approximation property (see e.g. [45, p. 30ff] for details on this concept). For a fixed $\varepsilon > 0$ there consequently exists a bounded linear operator $S : H \to H$ with $m := \operatorname{rank} S < \infty$ and

$$\|S\Phi(z) - \Phi(z)\|_H \leq \varepsilon, \quad z \in K.$$

Let $e_1, \ldots, e_m$ be an ONB of the image SH of H under S. Since $\langle e_j, S\Phi\rangle : Z \to \mathbb{R}, j = 1, \ldots, m$, are bounded measurable functions we then find that

$$\langle e_j, hS\Phi\rangle = h\langle e_j, S\Phi\rangle, \quad j = 1, \ldots, m,$$

are P-integrable. Consequently, they satisfy the limit relation (8), and by a well-known reformulation of almost sure convergence (see e.g. [43, Lem. 20.6]) hence there exists an n_ε such that with probability not less than $1 - \varepsilon$ we have both

$$\sup_{n\geq n_\varepsilon}\sup_{j=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \langle e_j, hS\Phi\rangle\circ Z_i(\omega) - \mathbb{E}_P\langle e_j, hS\Phi\rangle\right| \leq \varepsilon m^{-1/2}$$

and

$$\sup_{n\geq n_\varepsilon}\left|\frac{1}{n}\sum_{i=1}^n |h|\circ Z_i(\omega) - \mathbb{E}_P|h|\right| \leq \varepsilon.$$

Let us fix an $n \geq n_\varepsilon$ and an $\omega \in \Omega$ which satisfy these two inequalities. Using $h(z) = 0$ for all $z \in Z \setminus K$ we then have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n (h\Phi)\circ Z_i(\omega) - \frac{1}{n}\sum_{i=1}^n (hS\Phi)\circ Z_i(\omega)\right\|_H &\leq \frac{1}{n}\sum_{i=1}^n |h|\circ Z_i(\omega)\cdot\|\Phi\circ Z_i(\omega) - S\Phi\circ Z_i(\omega)\|_H\\ &\leq \frac{\varepsilon}{n}\sum_{i=1}^n |h|\circ Z_i(\omega)\\ &\leq \varepsilon\,(\varepsilon + \mathbb{E}_P|h|)\\ &\leq \varepsilon + \varepsilon\,\mathbb{E}_P|h|.\end{aligned}$$

Moreover, n and ω also satisfy

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n (hS\Phi)\circ Z_i(\omega) - \mathbb{E}_P hS\Phi\right\|_H &= \left(\sum_{j=1}^m \left|\left\langle e_j, \frac{1}{n}\sum_{i=1}^n (hS\Phi)\circ Z_i(\omega) - \mathbb{E}_P hS\Phi\right\rangle\right|^2\right)^{1/2}\\ &\leq \sqrt{m}\sup_{j=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \langle e_j, hS\Phi\rangle\circ Z_i(\omega) - \mathbb{E}_P\langle e_j, hS\Phi\rangle\right|\\ &\leq \varepsilon.\end{aligned}$$

In addition, $h(z) = 0$ for all $z \in Z \setminus K$ implies

$$\|\mathbb{E}_P hS\Phi - \mathbb{E}_P h\Phi\|_H \leq \int_K |h(z)|\cdot\|S\Phi(z) - \Phi(z)\|_H\,\mathrm{d}P(z) \leq \varepsilon\,\mathbb{E}_P|h|,$$

and consequently we can conclude

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n (h\Phi)\circ Z_i(\omega) - \mathbb{E}_P h\Phi\right\|_H &\leq \left\|\frac{1}{n}\sum_{i=1}^n (h\Phi)\circ Z_i(\omega) - \frac{1}{n}\sum_{i=1}^n (hS\Phi)\circ Z_i(\omega)\right\|_H\\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^n (hS\Phi)\circ Z_i(\omega) - \mathbb{E}_P hS\Phi\right\|_H + \|\mathbb{E}_P hS\Phi - \mathbb{E}_P h\Phi\|_H\\ &\leq 2\varepsilon\,(1 + \mathbb{E}_P|h|)\,.\end{aligned}$$

This shows

$$\mu\left(\left\{\omega\in\Omega : \sup_{n\geq n_\varepsilon}\left\|\frac{1}{n}\sum_{i=1}^{n}(h\Phi)\circ Z_i(\omega) - \mathbb{E}_P h\Phi\right\|_H \leq 2\varepsilon\,(1+\mathbb{E}_P|h|)\right\}\right)\geq 1-\varepsilon,$$

and hence [43, Lemma 20.6] yields the assertion for our special case.

Let us now prove the assertion for general $h\in L_1(P)$. To this end we may assume without loss of generality that $\|\Phi(z)\|\leq 1$ for all $z\in Z$. Let us fix an $\varepsilon>0$. Since Z is Polish the measures P and $|h|P$ are regular and hence there then exists a compact subset $K\subset Z$ with

$$P(Z\setminus K)\leq\varepsilon \quad\text{and}\quad \int_{Z\setminus K}|h|\,\mathrm{d}P\leq\varepsilon.$$

Now $g := \mathbf{1}_K h$ is a P-integrable function that vanishes outside the compact set K. Our preliminary considerations and the SLLN consequently show that there exists an $n_\varepsilon\geq 1$ such that with probability not less than $1-\varepsilon$ we have both

$$\sup_{n\geq n_\varepsilon}\left\|\frac{1}{n}\sum_{i=1}^{n}(g\Phi)\circ Z_i(\omega) - \mathbb{E}_P g\Phi\right\|_H\leq\varepsilon$$

and

$$\sup_{n\geq n_\varepsilon}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{Z\setminus K}|h|\right)\circ Z_i(\omega) - \mathbb{E}_P\mathbf{1}_{Z\setminus K}|h|\right|\leq\varepsilon.$$

Let us fix an $n\geq n_\varepsilon$ and an $\omega\in\Omega$ which satisfy these two inequalities. Using $h-g=\mathbf{1}_{Z\setminus K}h$ and $\|\Phi(z)\|\leq 1$ for all $z\in Z$ we then obtain

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}(h\Phi)\circ Z_i(\omega) - \mathbb{E}_P h\Phi\right\|_H &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}(h\Phi)\circ Z_i(\omega) - \frac{1}{n}\sum_{i=1}^{n}(g\Phi)\circ Z_i(\omega)\right\|_H \\
&\quad + \left\|\frac{1}{n}\sum_{i=1}^{n}(g\Phi)\circ Z_i(\omega) - \mathbb{E}_P g\Phi\right\|_H + \|\mathbb{E}_P g\Phi - \mathbb{E}_P h\Phi\|_H \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{Z\setminus K}|h|\right)\circ Z_i(\omega) + \varepsilon + \mathbb{E}_P\mathbf{1}_{Z\setminus K}|h| \\
&\leq \varepsilon + \mathbb{E}_P\mathbf{1}_{Z\setminus K}|h| + \varepsilon + \mathbb{E}_P\mathbf{1}_{Z\setminus K}|h| \\
&\leq 4\varepsilon.
\end{aligned}$$

Therefore we obtain

$$\mu\left(\left\{\omega\in\Omega : \sup_{n\geq n_\varepsilon}\left\|\frac{1}{n}\sum_{i=1}^{n}(h\Phi)\circ Z_i(\omega) - \mathbb{E}_P h\Phi\right\|_H \leq 4\varepsilon\right\}\right)\geq 1-\varepsilon,$$

and hence we obtain the assertion by another application of [43, Lemma 20.6].

Finally, if $\mathcal{Z}$ only satisfies the WLLN then we obtain the assertion by omitting the terms $\sup_{n\geq n_\varepsilon}$ in the above proof. Moreover, for processes satisfying only a law of large numbers for events we have to use Lemma 2.5 instead of Definition 2.6. ■

In order to prove Theorem 2.17 we finally need the following technical lemma:

Lemma 4.5. *Let $F:(0,\infty)\times\mathbb{N}\to[0,\infty)$ be a function with $\lim_{n\to\infty}F(\lambda,n)=0$ for all $\lambda>0$. Then there exists a sequence $(\lambda_n)\subset(0,1]$ with*

$$\lim_{n\to\infty}\lambda_n = 0$$

and

$$\lim_{n\to\infty}F(\lambda_n,n)=0.$$

Proof. For $k\geq 1$ there exists an $n_k\geq 1$ such that for all $n\geq n_k$ we have

$$F(k^{-1},n) < k^{-1}. \tag{41}$$

Obviously, we may assume without loss of generality that $n_k<n_{k+1}$ for all $k\geq 1$. For $n\geq 1$ we write

$$\lambda_n := \begin{cases} 1 & \text{if } 1\leq n<n_1 \\ k^{-1} & \text{if } n_k\leq n<n_{k+1}. \end{cases}$$

Now let $\varepsilon>0$. Then there exists an integer $k\geq 1$ with $k^{-1}\leq\varepsilon$. Let us fix an $n\geq n_k$. Then there exists an $i\geq k$ with

$n_i \le n < n_{i+1}$, and consequently we have $\lambda_n = i^{-1}$. This gives

$$\lambda_n = i^{-1} \le k^{-1} \le \varepsilon,$$

and since (41) together with $n_i \le n$ yields $F(i^{-1}, n) \le i^{-1}$ we also find

$$F(\lambda_n, n) = F(i^{-1}, n) \le i^{-1} \le \varepsilon.$$

These estimates show the assertion. ■

Proof of Theorem 2.17. We only show the assertion in the case of $\mathcal{Z}$ satisfying the SLLNE. Since L is locally bounded, the function $L(.,.,0)$ is bounded and hence we may assume without loss of generality that $\mathcal{R}_{L,Q}(0) \le 1$ for all distributions Q on $X \times Y$. Let us fix a distribution Q on $X \times Y$ and a $\lambda > 0$. Since $f_{Q,\lambda}$ is a minimizer of the regularized risk defined by Q we then have

$$\lambda \|f_{Q,\lambda}\|_H^2 \le \lambda \|f_{Q,\lambda}\|_H^2 + \mathcal{R}_{L,Q}(f_{Q,\lambda}) \le \mathcal{R}_{L,Q}(0) \le 1$$

and hence we conclude $\|f_{Q,\lambda}\|_H \le \lambda^{-1/2}$ for *all* distributions Q on $X \times Y$ and all $\lambda > 0$. Moreover, we may assume without loss of generality that $\|k\|_\infty \le 1$, so that we have $\|f\|_\infty \le \|f\|_H$ for all $f \in H$. Now, let us fix an $\varepsilon > 0$. Since a simple argument shows that $\lim_{\lambda\to 0} \mathcal{R}_{L,P}(f_{P,\lambda}) = \mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P}$ we then find

$$\begin{aligned}
\left|\mathcal{R}_{L,P}(f_{T_n(\omega),\lambda}) - \mathcal{R}^*_{L,P}\right| &\le \left|\mathcal{R}_{L,P}(f_{T_n(\omega),\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda})\right| + \left|\mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P}\right| \\
&\le |L|_{\lambda^{-1/2},1} \|f_{T_n(\omega),\lambda} - f_{P,\lambda}\|_\infty + \varepsilon \\
&\le \frac{|L|_{\lambda^{-1/2},1}}{\lambda} \left\|\mathbb{E}_{T_n(\omega)} h_\lambda \Phi - \mathbb{E}_P h_\lambda \Phi\right\|_H + \varepsilon
\end{aligned}$$

for all $n \ge 1$, $\omega \in \Omega$, and all sufficiently small $\lambda > 0$, where $h_\lambda : X \times Y \to \mathbb{R}$ is the function according to Theorem 4.2, and $\mathbb{E}_{T_n(\omega)}$ denotes the expectation operator with respect to the empirical distribution associated to the training set $T_n(\omega) = ((X_1(\omega), Y_1(\omega)), \ldots, (X_n(\omega), Y_n(\omega)))$, i.e. $\mathbb{E}_{T_n(\omega)} g = \frac{1}{n}\sum_{i=1}^n g(X_i(\omega), Y_i(\omega))$. Furthermore, for all $\lambda \in (0, \varepsilon]$ and $n \ge 1$ we have

$$\begin{aligned}
&\mu\left(\left\{\omega \in \Omega : \sup_{m\ge n} \frac{|L|_{\lambda^{-1/2},1}}{\lambda} \left\|\mathbb{E}_{T_m(\omega)} h_\lambda \Phi - \mathbb{E}_P h_\lambda \Phi\right\|_H \ge \varepsilon\right\}\right) \\
&\le \mu\left(\left\{\omega \in \Omega : \sup_{m\ge n} \left\|\mathbb{E}_{T_m(\omega)} h_\lambda \Phi - \mathbb{E}_P h_\lambda \Phi\right\|_H \ge \frac{\lambda^2}{|L|_{\lambda^{-1/2},1}}\right\}\right) \\
&=: F(\lambda, n).
\end{aligned}$$

Moreover, by Theorem 4.2 we know that h_λ is a bounded function for all $\lambda > 0$ and consequently, Lemma 4.4 yields $\lim_{n\to\infty} F(\lambda, n) = 0$ for all $\lambda \in (0, \varepsilon]$. Now Lemma 4.5 shows that there exists a sequence (λ_n) with $\lambda_n \to 0$ and $F(\lambda_n, n) \to 0$. For fixed $\delta > 0$ there consequently exists an $n_0 \ge 1$ such that for all $n \ge n_0$ we have $|\mathcal{R}_{L,P}(f_{P,\lambda_n}) - \mathcal{R}^*_{L,P}| \le \varepsilon$, $\lambda_n \le \varepsilon$, and $F(\lambda_n, n) \le \delta$. For such n our previous considerations then show

$$\begin{aligned}
&\mu\left(\left\{\omega \in \Omega : \sup_{m\ge n} \left|\mathcal{R}_{L,P}(f_{T_m(\omega),\lambda_m}) - \mathcal{R}^*_{L,P}\right| \ge 2\varepsilon\right\}\right) \\
&\le \mu\left(\left\{\omega \in \Omega : \sup_{m\ge n} \frac{|L|_{\lambda_m^{-1/2},1}}{\lambda_m} \left\|\mathbb{E}_{T_m(\omega)} h_{\lambda_m} \Phi - \mathbb{E}_P h_{\lambda_m} \Phi\right\|_H \ge \varepsilon\right\}\right) \\
&\le F(\lambda_n, n) \\
&\le \delta.
\end{aligned}$$

This shows the assertion. ■

Proof of Theorem 2.18. Again, we only show the assertion in the case of $\mathcal{Z}$ satisfying the SLLN. Obviously, we may assume without loss of generality that $\|k\|_\infty \le 1$, so that we have $\|f\|_\infty \le \|f\|_H$ for all $f \in H$. Moreover, since $|P|_p < \infty$ we may additionally assume without loss of generality that both $|P|_p \le 1$ and $\mathcal{R}_{L,P}(0) \le 1$. Note that the latter assumption immediately yields

$$\|f_{P,\lambda}\|_H \le \lambda^{-1/2}$$

for all $\lambda > 0$. Let $\psi : \mathbb{R} \to [0, \infty)$ be the function satisfying $L(y, t) = \psi(y - t)$, $y, t \in \mathbb{R}$. The assumption $|P|_p < \infty$ then guarantees $\psi \in L_1(P)$ and hence the SLLN shows

$$\lim_{n\to\infty} \mathcal{R}_{L,T_n(\omega)}(0) = \lim_{n\to\infty} \mathbb{E}_{T_n(\omega)} \psi = \mathbb{E}_P \psi = \mathcal{R}_{L,P}(0) \tag{42}$$

for μ-almost all $\omega \in \Omega$. Moreover, we have $\lambda \|f_{T_n(\omega),\lambda}\|_H^2 \le \mathcal{R}_{L,T_n(\omega)}(0)$ for all $n \ge 1$, $\lambda > 0$, and $\omega \in \Omega$. Consequently the "local Lipschitz continuity" of the L-risk which follows from (13) as shown in [4, Lemma 20] together with Theorem 4.3 yields

$$\begin{aligned}\left|\mathcal{R}_{L,P}(f_{T_n(\omega),\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda})\right| &\le c_p \left(|P|_{p-1}^{p-1} + \|f_{T_n(\omega),\lambda}\|_\infty^{p-1} + \|f_{P,\lambda}\|_\infty^{p-1} + 1\right) \|f_{T_n(\omega),\lambda} - f_{P,\lambda}\|_\infty \\ &\le \frac{c_p}{\lambda}\left(2 + \left(\frac{\mathcal{R}_{L,T_n(\omega)}(0)}{\lambda}\right)^{\frac{p-1}{2}} + \lambda^{-\frac{p-1}{2}}\right) \|\mathbb{E}_{T_n(\omega)} h_\lambda \Phi - \mathbb{E}_P h_\lambda \Phi\|_H\end{aligned}$$

for all $n \ge 1$, $\lambda > 0$, and $\omega \in \Omega$. Let us fix an $\varepsilon > 0$. For $\lambda \in (0, \varepsilon]$ and $n \ge 1$ we then obtain

$$\begin{aligned}&\mu\left(\left\{\omega \in \Omega : \sup_{m\ge n} \left|\mathcal{R}_{L,P}(f_{T_m(\omega),\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda})\right| \ge \varepsilon\right\}\right) \\ &\le \mu\left(\left\{\omega \in \Omega : \sup_{m\ge n}\left(2 + \left(\frac{\mathcal{R}_{L,T_m(\omega)}(0)}{\lambda}\right)^{\frac{p-1}{2}} + \lambda^{-\frac{p-1}{2}}\right) \|\mathbb{E}_{T_m(\omega)} h_\lambda \Phi - \mathbb{E}_P h_\lambda \Phi\|_H \ge \frac{\lambda^2}{c_p}\right\}\right) \\ &=: F(\lambda, n).\end{aligned}$$

Moreover, Theorem 4.3 ensures $h_\lambda \in L_1(P)$ for all $\lambda > 0$ and hence Lemma 4.4 together with (42) shows that $\lim_{n\to\infty} F(\lambda, n) = 0$ for all $\lambda \in (0, \varepsilon]$. Now the rest of the proof is analogous to the proof of Theorem 2.17. ■

4.4. Proofs from Section 3.1

Proof of Proposition 3.2. Obviously, it suffices to show that being AMS implies the WLLNE. To this end let P be the stationary mean of $(\mathcal{Z}, \mu)$. Then there exists an $n_0 \ge 1$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu \mathbf{1}_B \circ Z_i - P(B)\right| < \frac{\varepsilon}{2}, \quad n \ge n_0,$$

and hence Markov's inequality yields

$$\begin{aligned}&\mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_B \circ Z_i(\omega) - P(B)\right| \ge \varepsilon\right\}\right) \\ &\le \mu\left(\left\{\omega \in \Omega : \left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_B \circ Z_i(\omega) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu \mathbf{1}_B \circ Z_i\right| \ge \frac{\varepsilon}{2}\right\}\right) \\ &\le 4\varepsilon^{-2} n^{-2} \mathbb{E}_\mu \left(\sum_{i=1}^n \left(\mathbf{1}_B \circ Z_i - \mathbb{E}_\mu \mathbf{1}_B \circ Z_i\right)\right)^2\end{aligned}$$

for all $n \ge n_0$. Let us write $h_i := \mathbf{1}_B \circ Z_i - \mathbb{E}_\mu \mathbf{1}_B \circ Z_i$, $i \ge 1$. Then we have $\mathbb{E}_\mu h_i = 0$ and $h_i(\omega) \in [-1, 1]$ for all $i \ge 1$ and all $\omega \in \Omega$. Consequently, (20) gives $R_\infty^{\mathbb{E}}(\mathcal{Z}, \mu, i, j) \le 2\pi\alpha(\mathcal{Z}, \mu, i, j)$, $i, j \ge 1$, and hence we obtain

$$\mathbb{E}_\mu \left(\sum_{i=1}^n \left(\mathbf{1}_B \circ Z_i - \mathbb{E}_\mu \mathbf{1}_B \circ Z_i\right)\right)^2 = \mathbb{E}_\mu \sum_{i=1}^n h_i^2 + 2\mathbb{E}_\mu \sum_{i=1}^n \sum_{j=1}^{i-1} h_i h_j \le n + 4\pi \sum_{i=1}^n \sum_{j=1}^{i-1} \alpha(\mathcal{Z}, \mu, i, j).$$

Combining the estimates then yields the assertion. ■

4.5. Proofs from Section 3.2

Proof of Theorem 3.3. Let $\mathcal{B}$ be the σ-algebra of Z. We write $P_n(B) := \frac{1}{n}\sum_{i=1}^n \mu(Z_i \in B)$ for $B \in \mathcal{B}$ and $n \ge 1$. Then P_n is obviously a probability measure on $\mathcal{B}$ for all $n \ge 1$. Let us first show that

$$\lim_{n\to\infty} \mathcal{R}_{L,P}(f_{P_n,\lambda_n}) = \mathcal{R}_{L,P}^*. \tag{43}$$

To this end we first observe that the assumption (24) yields

$$\begin{aligned}\mathcal{R}_{L,P}(f_{P_n,\lambda_n}) &\le \lambda_n \|f_{P_n,\lambda_n}\|_H^2 + \mathcal{R}_{L,P_n}(f_{P_n,\lambda_n}) + C\|L \circ f_{P_n,\lambda_n}\|_\infty n^{-\alpha} \\ &\le \lambda_n \|f_{P,\lambda_n}\|_H^2 + \mathcal{R}_{L,P_n}(f_{P,\lambda_n}) + C\|L \circ f_{P_n,\lambda_n}\|_\infty n^{-\alpha} \\ &\le \lambda_n \|f_{P,\lambda_n}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda_n}) + Cn^{-\alpha}\left(\|L \circ f_{P,\lambda_n}\|_\infty + \|L \circ f_{P_n,\lambda_n}\|_\infty\right)\end{aligned} \tag{44}$$

for all $n \ge 1$. Now $\mathcal{R}_{L,P,H}^* = \mathcal{R}_{L,P}^*$ together with $\lambda_n \to 0$ yields $\lambda_n \|f_{P,\lambda_n}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda_n}) \to \mathcal{R}_{L,P}^*$. Moreover, for every distribution Q on Z we have

$$\|L \circ f_{Q,\lambda}\|_\infty \le c + |L|_{\|f_{Q,\lambda}\|_\infty,1}\|f_{Q,\lambda}\|_\infty \le c + |L|_{B_\lambda,1} B_\lambda$$

by (11) and Theorem 4.1. In addition, $(|L|_{B_{\lambda_n},1})$ is a non-decreasing sequence and the sequence (B_{λ_n}) is dominated by the sequence $(\lambda_n^{-1/2})$. Consequently, (26) implies $n^{-\alpha}|L|_{B_{\lambda_n},1} B_{\lambda_n} \to 0$ and hence we find (43). Let us now fix an $\varepsilon > 0$. Then Theorem 4.2 and Markov's inequality yield

$$\begin{aligned}
&\mu\left(\{\omega\in\Omega : \left|\mathcal{R}_{L,P}(f_{T_n(\omega),\lambda_n}) - \mathcal{R}_{L,P}(f_{P_n,\lambda_n})\right| \geq \varepsilon\}\right)\\
&\quad\leq \mu\left(\left\{\omega\in\Omega : |L|_{B_{\lambda_n},1}\, \|f_{T_n(\omega),\lambda_n} - f_{P_n,\lambda_n}\|_\infty \geq \varepsilon\right\}\right)\\
&\quad\leq \mu\left(\left\{\omega\in\Omega : \|k\|_\infty |L|_{B_{\lambda_n},1}\, \left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H \geq \varepsilon\lambda_n\right\}\right)\\
&\quad\leq \frac{\|k\|_\infty^2 |L|_{B_{\lambda_n},1}^2}{\varepsilon^2\lambda_n^2}\,\mathbb{E}_{\omega\sim\mu}\left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H^2
\end{aligned}$$

where h_n is the function according to Theorem 4.2 for the distribution P_n and the regularization parameter λ_n. Let us define

$$g_{n,i} := (h_n\Phi)\circ(X_i, Y_i) - \mathbb{E}_\mu(h_n\Phi)\circ(X_i, Y_i)$$

for $n \geq 1$ and $i = 1, \ldots, n$. Then we have $\mathbb{E}_\mu g_{n,i} = 0$ and Theorem 4.2 yields

$$\|g_{n,i}\|_\infty \leq 2\sup_{\omega\in\Omega}\|(h_n\Phi)\circ(X_i, Y_i)(\omega)\|_H \leq 2\,\|h_n\|_\infty\,\|k\|_\infty \leq 2\,\|k\|_\infty |L|_{B_{\lambda_n},1}.$$

Consequently, (20) and (19) show that there exists a universal constant $c \geq 1$ such that

$$\begin{aligned}
\mathbb{E}_{\omega\sim\mu}\left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H^2 &= n^{-2}\,\mathbb{E}_{\omega\sim\mu}\left\|\sum_{i=1}^n (h_n\Phi)\circ(X_i, Y_i)(\omega) - \mathbb{E}_\mu(h_n\Phi)\circ(X_i, Y_i)\right\|_H^2\\
&= n^{-2}\sum_{i=1}^n \mathbb{E}_\mu\langle g_{n,i}, g_{n,i}\rangle + 2n^{-2}\sum_{i=1}^n\sum_{j=1}^{i-1}\mathbb{E}_\mu\langle g_{n,i}, g_{n,j}\rangle\\
&\leq n^{-2}\sum_{i=1}^n \|g_{n,i}\|_\infty^2 + 2n^{-2}\sum_{i=1}^n\sum_{j=1}^{i-1} R_\infty^H(\mathcal{Z},\mu,i,j)\|g_{n,i}\|_\infty\|g_{n,j}\|_\infty\\
&\leq 4n^{-1}\|k\|_\infty^2|L|_{B_{\lambda_n},1}^2 + c\,\|k\|_\infty^2|L|_{B_{\lambda_n},1}^2 n^{-2}\sum_{i=1}^n\sum_{j=1}^{i-1}\alpha(\mathcal{Z},\mu,i,j)
\end{aligned}$$

for all $n \geq 1$. By combining all estimates and using (26) we then obtain the assertion. ■

Proof of Theorem 3.4. Without loss of generality we assume that $\|k\|_\infty \leq 1$ and $|\mu_{(X_i,Y_i)}|_q \leq 1$ for all $i \geq 1$. In addition, we can obviously, also assume that $\lambda_n \in (0, 1]$ for all $n \geq 1$. Now, we define $P_n(B) := \frac{1}{n}\sum_{i=1}^n \mu(Z_i \in B)$ for measurable $B \subset X\times\mathbb{R}$ and $n \geq 1$. For $r \in [1, q]$ a simple calculation then shows

$$|P_n|_r^r = \int_{X\times\mathbb{R}}|y|^r\mathrm{d}P_n(x,y) = \frac{1}{n}\sum_{i=1}^n\int_{X\times\mathbb{R}}|y|^r\mathrm{d}\mu_{(X_i,Y_i)}(x,y) = \frac{1}{n}\sum_{i=1}^n|\mu_{(X_i,Y_i)}|_r^r \leq 1. \tag{45}$$

Moreover, [43, Thm. 23.8] together with Fatou's lemma yields

$$\begin{aligned}
|P|_r^r = \int_0^\infty P\left(\{(x,y)\in X\times\mathbb{R} : |y|^r \geq t\}\right)\mathrm{d}t &= \int_0^\infty \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mu\left(\{\omega\in\Omega : |Y_i(\omega)|^r \geq t\}\right)\mathrm{d}t\\
&\leq \liminf_{n\to\infty}\int_0^\infty \frac{1}{n}\sum_{i=1}^n \mu\left(\{\omega\in\Omega : |Y_i(\omega)|^r \geq t\}\right)\mathrm{d}t\\
&\leq \liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^n |\mu_{(X_i,Y_i)}|_r^r\\
&\leq 1.
\end{aligned}$$

Having finished these preparations we can now begin with the actual proof. To this end first observe that we obtain

$$\mathcal{R}_{L,P}(f_{P_n,\lambda_n}) \leq \lambda_n\|f_{P,\lambda_n}\|_H^2 + \mathcal{R}_{L,P}(f_{P,\lambda_n}) + Cn^{-\alpha}\left(\|L\circ f_{P,\lambda_n}\|_{L_1(P)} + \|L\circ f_{P_n,\lambda_n}\|_{L_1(P)}\right)$$

as in (44). Moreover, we obviously have $\|L\circ f_{P,\lambda_n}\|_{L_1(P)} = \mathcal{R}_{L,P}(f_{P,\lambda_n}) \leq \mathcal{R}_{L,P}(0) \leq c$ for some constant c independent of n. In addition, (45) yields

$$\begin{aligned}
\|L\circ f_{P_n,\lambda_n}\|_{L_1(P)} &= \int_{X\times Y}\psi\left(y - f_{P_n,\lambda_n}(x)\right)\mathrm{d}P(x,y)\\
&\leq \tilde{c}_p\int_{X\times Y} 1 + |y|^p + |f_{P_n,\lambda_n}(x)|^p\mathrm{d}P(x,y)\\
&\leq 2\tilde{c}_p + \tilde{c}_p\|f_{P_n,\lambda_n}\|_\infty^p\\
&\leq 2\tilde{c}_p + \tilde{c}_p\|k\|_\infty^p\left(\frac{\mathcal{R}_{L,P_n}(0)}{\lambda_n}\right)^{\frac{p}{2}}\\
&\leq 2c_p + c_p\lambda_n^{-\frac{p}{2}},
\end{aligned}$$

where $\tilde{c}_p$ and c_p are constants only depending on L and p. Combining these estimates with $\lim_{\lambda\to 0}\mathcal{R}_{L,P}(f_{P,\lambda}) = \mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P}$ and (30) we then obtain $\lim_{n\to\infty}\mathcal{R}_{L,P}(f_{P_n,\lambda_n}) = \mathcal{R}^*_{L,P}$.

Now let us assume that we have an $\omega \in \Omega$ and an $n \geq 1$ with $\|f_{T_n(\omega),\lambda_n} - f_{P_n,\lambda_n}\|_H \leq 1$. For $p > 1$ the "local Lipschitz continuity" of the L-risk which follows from (13) as shown in [4, Lemma 20] together with $\lambda_n \leq 1$ then yields

$$\begin{aligned}
|\mathcal{R}_{L,P}(f_{P_n,\lambda_n}) - \mathcal{R}_{L,P}(f_{T_n(\omega),\lambda_n})| &\leq C_p\left(|P|_{p-1}^{p-1} + \|f_{P_n,\lambda_n}\|_\infty^{p-1} + \|f_{T_n(\omega),\lambda_n}\|_\infty^{p-1} + 1\right)\|f_{P_n,\lambda_n} - f_{T_n(\omega),\lambda_n}\|_\infty \\
&\leq C_p\left(2 + 2\|f_{P_n,\lambda_n}\|_\infty^{p-1} + \|f_{T_n(\omega),\lambda_n} - f_{P_n,\lambda_n}\|_\infty^{p-1}\right)\|f_{P_n,\lambda_n} - f_{T_n(\omega),\lambda_n}\|_H \\
&\leq C_p\left(3 + 2\left(\frac{\mathcal{R}_{L,P_n}(0)}{\lambda_n}\right)^{\frac{p-1}{2}}\right)\|f_{P_n,\lambda_n} - f_{T_n(\omega),\lambda_n}\|_H \\
&\leq \bar{C}_p\,\lambda_n^{-\frac{p-1}{2}}\|f_{P_n,\lambda_n} - f_{T_n(\omega),\lambda_n}\|_H \\
&\leq \bar{C}_p\,\lambda_n^{-\frac{p+1}{2}}\left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H,
\end{aligned}$$

where $C_p \geq 1$ and $\bar{C}_p \geq 1$ are constants only depending on p and L, and h_n is the function according to Theorem 4.2 for the distribution P_n and the regularization parameter λ_n. Moreover, for $p = 1$ we see that L is Lipschitz continuous by [4, Lemma 4] and hence the above estimate is also true in this case. Let us now define

$$g_{n,i} := (h_n\Phi)\circ(X_i, Y_i) - \mathbb{E}_\mu(h_n\Phi)\circ(X_i, Y_i)$$

for $n \geq 1$ and $i = 1, \ldots, n$. Then we have $\mathbb{E}_\mu g_{n,i} = 0$ and for $s := \frac{q}{p-1}$ we find

$$\begin{aligned}
\|g_{n,i}\|_{L_s(\mu)} \leq 2\|h_n\|_{L_s(\mu_{(X_i,Y_i)})} &\leq 128 c_L\left(1 + |\mu_{(X_i,Y_i)}|_q^{p-1} + \|f_{P_n,\lambda_n}\|_\infty^{p-1}\right) \\
&\leq 128 c_L\left(2 + \left(\frac{\mathcal{R}_{L,P_n}(0)}{\lambda_n}\right)^{\frac{p-1}{2}}\right) \\
&\leq C_{L,p}\,\lambda_n^{-\frac{p-1}{2}},
\end{aligned}$$

where $C_{L,p} > 0$ is a constant only depending on L and p. For $\delta > 0$ Markov's inequality together with $s \geq 2$, (20) and (19) thus yields

$$\begin{aligned}
&\mu\left(\{\omega \in \Omega : \left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H \geq \delta\}\right) \\
&\leq \frac{1}{\delta^2 n^2}\left(\sum_{i=1}^n \mathbb{E}_\mu\langle g_{n,i}, g_{n,i}\rangle + 2\sum_{i=1}^n\sum_{j=1}^{i-1}\mathbb{E}_\mu\langle g_{n,i}, g_{n,j}\rangle\right) \\
&\leq \frac{1}{\delta^2 n^2}\left(\sum_{i=1}^n \|g_{n,i}\|_{L_s(\mu)}^2 + 2\sum_{i=1}^n\sum_{j=1}^{i-1} R_s^H(\mathcal{Z}, \mu, i, j)\|g_{n,i}\|_{L_s(\mu)}\|g_{n,j}\|_{L_s(\mu)}\right) \\
&\leq \frac{\bar{C}_{L,p}}{\delta^2\lambda_n^{p-1}n} + \frac{\bar{C}_{L,p}}{\delta^2\lambda_n^{p-1}n^2}\sum_{i=1}^n\sum_{j=1}^{i-1}\alpha^{1-\frac{2p-2}{q}}(\mathcal{Z}, \mu, i, j)\varphi_{\mathrm{sym}}^{\frac{2p-2}{q}}(\mathcal{Z}, \mu, i, j) \\
&\leq \frac{(1+C)\bar{C}_{L,p}}{\delta^2\lambda_n^{p-1}n^\beta},
\end{aligned}$$

where $\bar{C}_{L,p} > 0$ is another constant only depending on L and p. Let us now fix an $\varepsilon \in (0, 1]$. For $\omega \in \Omega$ and $n \geq 1$ with

$$\left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H < \frac{\varepsilon\lambda_n^{(p+1)/2}}{\bar{C}_p}$$

we then have $\|f_{T_n(\omega),\lambda_n} - f_{P_n,\lambda_n}\|_H < \frac{\varepsilon\lambda_n^{(p-1)/2}}{\bar{C}_p} \leq 1$, and consequently we can conclude

$$\begin{aligned}
&\mu\left(\{\omega \in \Omega : |\mathcal{R}_{L,P}(f_{P_n,\lambda_n}) - \mathcal{R}_{L,P}(f_{T_n(\omega),\lambda_n})| < \varepsilon\}\right) \\
&\geq \mu\left(\left\{\omega \in \Omega : \left\|\mathbb{E}_{T_n(\omega)}h_n\Phi - \mathbb{E}_{P_n}h_n\Phi\right\|_H < \frac{\varepsilon\lambda_n^{(p+1)/2}}{\bar{C}_p}\right\}\right) \\
&\geq 1 - \frac{(1+C)\bar{C}_{L,p}\bar{C}_p^2}{\varepsilon^2\lambda_n^{2p}n^\beta}.
\end{aligned}$$

Using (31) then yields the assertion. ■

References

[1] I. Steinwart, Support vector machines are universally consistent, J. Complexity 18 (2002) 768–791.
[2] T. Zhang, Statistical behaviour and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–134.
[3] I. Steinwart, Consistency of support vector machines and other regularized kernel machines, IEEE Trans. Inform. Theory 51 (2005) 128–142.
[4] A. Christmann, I. Steinwart, Consistency and robustness of kernel based regression, Bernoulli 13 (2007) 799–819.
[5] D.R. Chen, Q. Wu, Y.M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[6] I. Steinwart, C. Scovel, Fast rates for support vector machines, in: Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005, Springer, 2005, pp. 279–294.
[7] G. Blanchard, O. Bousquet, P. Massart, Statistical performance of support vector machines, Ann. Statist. 36 (2008) 489–531.
[8] V. Koltchinskii, O. Beznosova, Exponential convergence rates in classification, in: Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005, Springer, 2005, pp. 295–307.
[9] I. Steinwart, C. Scovel, Fast rates for support vector machines using Gaussian kernels, Ann. Statist. 35 (2007) 575–607.
[10] A.B. Nobel, Limits to classification and regression estimation from ergodic processes, Ann. Statist. 27 (1999) 262–273.
[11] M. Vidyasagar, A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems, 2nd edition, Springer, London, 2002.
[12] A. Irle, On consistency in nonparametric estimation under mixing conditions, J. Multivariate Anal. 60 (1997) 123–147.
[13] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, New York, 2002.
[14] D.S. Modha, E. Masry, Memory-universal prediction of stationary random processes, IEEE Trans. Inform. Theory 44 (1998) 117–133.
[15] R. Meir, Nonparametric time series prediction through adaptive model selection, Mach. Learn. 39 (2000) 5–34.
[16] L. Györfi, W. Härdle, P. Sarda, P. Vieu, Nonparametric Curve Estimation from Time Series, Springer, Berlin, 1989.
[17] D. Bosq, Nonparametric Statistics for Stochastic Processes, 2nd edition, Springer, New York, 1998.
[18] R.M. Gray, J.C Kieffer, Asymptotically mean stationary measures, Ann. Probab. 8 (1980) 962–973.
[19] P. Révész, The Laws of Large Numbers, Academic Press, New York, 1968.
[20] U. Krengel, Ergodic Theorems, de Gruyter, Berlin, 1985.
[21] K. Petersen, Ergodic Theory, paperback edition, Cambridge University Press, 1989.
[22] J.L. Doob, Stochastic Processes, Wiley, New York, 1953.
[23] R. Bhattacharya, E.C. Waymire, Iterated random maps and some classes of markov processes, in: D.N. Shanbhag, C.R. Rao (Eds.), in: Handbook of Statistics, vol. 19, North-Holland, 2001, pp. 145–170.
[24] J.R. Norris, Markov Chains, Cambridge University Press, 1997.
[25] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[26] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[27] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, 2002.
[28] I. Steinwart, D. Hush, C. Scovel, Function classes that approximate the Bayes risk, in: Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006, Springer, 2006, pp. 79–93.
[29] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, J. Mach. Learn. Res. 2 (2001) 67–93.
[30] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[31] R.C. Bradley, Basic properties of strong mixing conditions. A survey and some open questions, Probab. Surveys 2 (2005) 107–144.
[32] R.C. Bradley, W. Bryc, S. Janson, Remarks on the foundations of measures of dependence, in: M.L. Puri, J.P. Vilaplana, W. Wertz (Eds.), New Perspectives in Theoretical and Applied Statistics, Wiley, 1987, pp. 421–437.
[33] H. Dehling, W. Philipp, Almost sure invariance principles for weakly dependent vector-valued random variables, Ann. Probab. 10 (1982) 689–701.
[34] R.C. Bradley, Introduction to strong mixing conditions, volume 1, Technical Report, Department of Mathematics, Indiana University, Bloomington, Custom Publishing of I.U., Bloomington, 2005.
[35] J. Fan, Q. Yao, Nonlinear Time Series, Springer, New York, 2003.
[36] R.C. Bradley, Introduction to strong mixing conditions, volume 2, Technical Report, Department of Mathematics, Indiana University, Bloomington, Custom Publishing of I.U., Bloomington, 2005.
[37] R.C. Bradley, Introduction to strong mixing conditions, volume 3, Technical Report, Department of Mathematics, Indiana University, Bloomington, Custom Publishing of I.U., Bloomington, 2005.
[38] L. Zhengyan, L. Chuanrong, Limit Theory for Mixing Dependent Random Variables, Science Press and Kluwer, New York, Dordrecht, 1996.
[39] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[40] I. Steinwart, D. Hush, C. Scovel, A new concentration result for regularized risk minimizers, in: E. Giné, V. Koltchinskii, W. Li, J. Zinn (Eds.), High Dimensional Probability IV, in: Lecture Notes–Monograph Series, vol. 51, Institute of Mathematical Statistics, Beachwood, OH, 2006, pp. 260–275.
[41] A.C. Lozano, S.R. Kulkarni, R.E. Schapire, Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006.
[42] N. Dunford, J.T. Schwartz, Linear Operators, Part I: General Theory, Wiley Classics Library edition, Wiley, New York, 1988.
[43] H. Bauer, Measure and Integration Theory, De Gruyter, Berlin, 2001.
[44] E. DeVito, L. Rosasco, A. Caponnetto, M. Piana, A. Verri, Some properties of regularized kernel methods, J. Mach. Learn. Res. 5 (2004) 1363–1390.
[45] J. Lindenstrauss, L. Tzafriri, Classical Banach Spaces I, Springer, Berlin, 1977.